Introducción a ML y GenAI

Árboles de Decisión para Regresión

Ariel Ramos Vela

03-10-2024

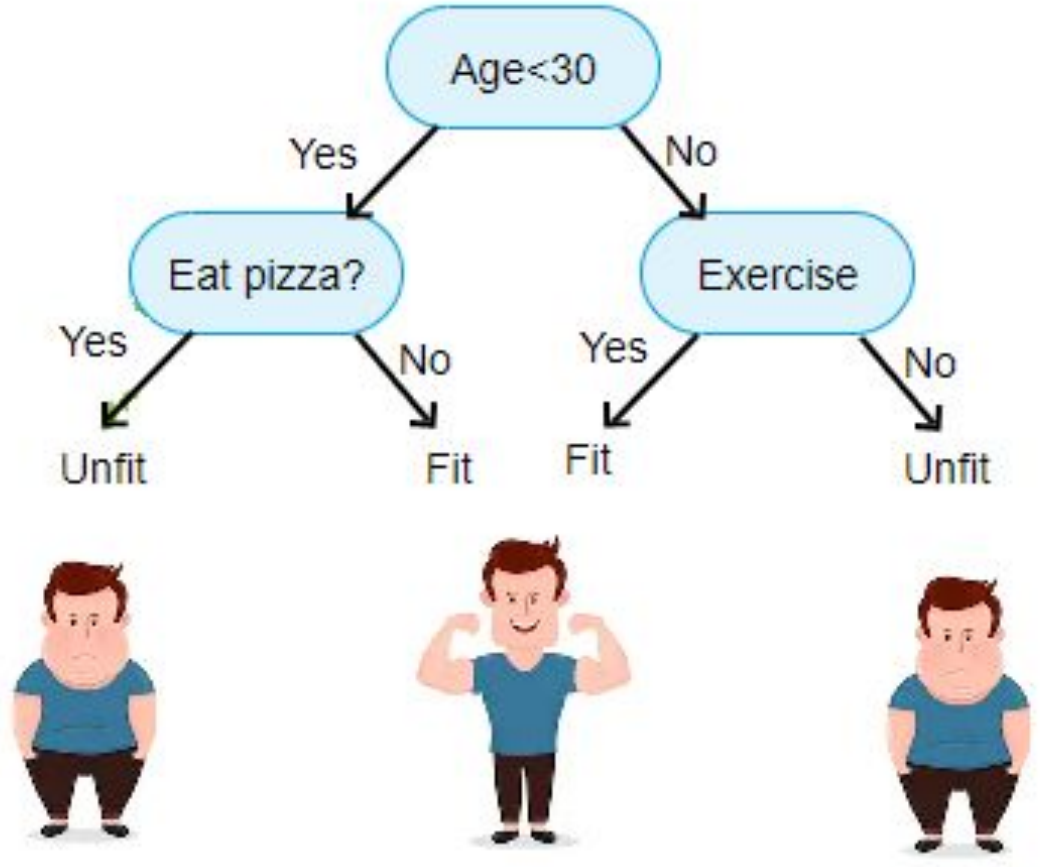
Agenda

1. Recapitulación: Árboles de decision
2. ¿Cómo utilizarlos para regression?
3. Construcción del árbol de decision
4. Ventajas y desventajas
5. Conclusiones
6. Taller 7

Recapitulación: ¿Qué es un árbol de decision (Decision Trees)?

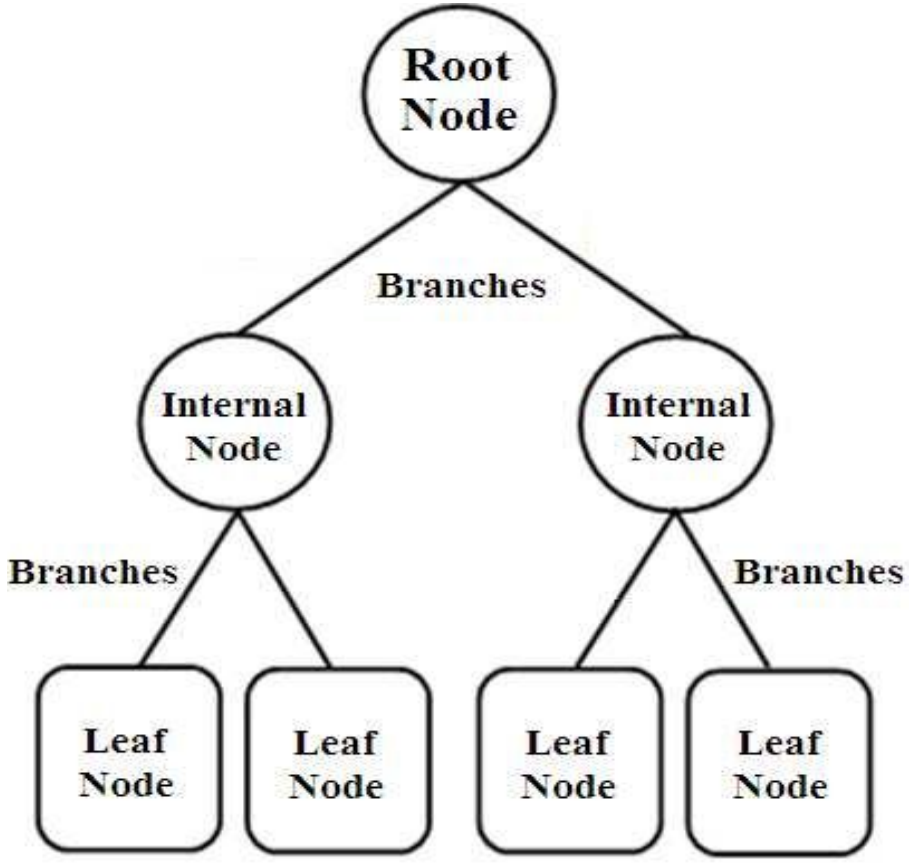
Definición: Un árbol de decisión es un modelo predictivo que divide iterativamente los datos en subconjuntos basados en características específicas.

Se utiliza tanto para clasificación como para **regresión**.



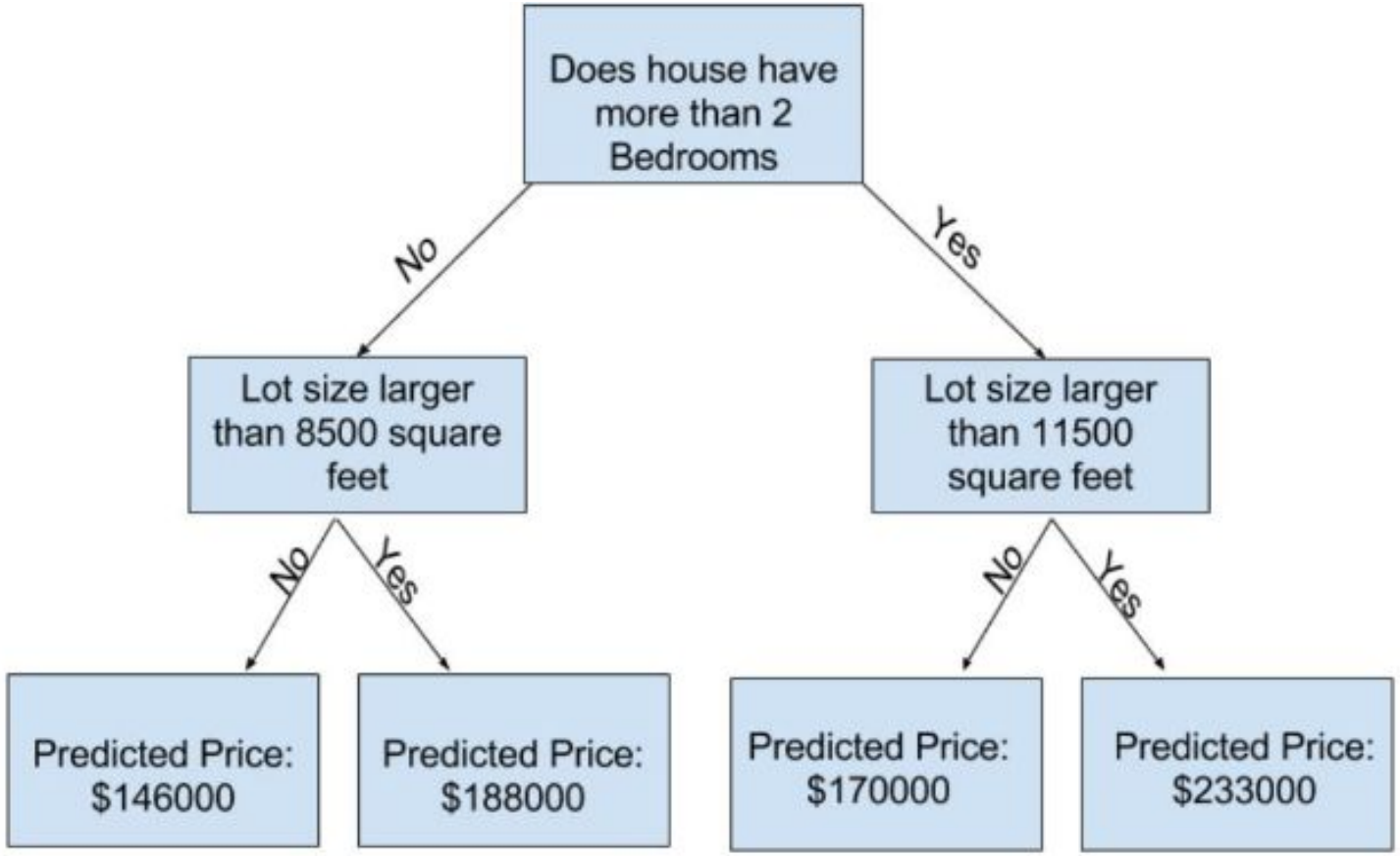
Recapitulación: Componentes de un árbol de decisión

- **Nodos**: Preguntas o decisiones basadas en atributos.
- **Ramas (branches)**: Resultado de una decisión.
- **Hojas (leaf)**: Resultados finales o clases.
- **Raíz (root)**: Nodo inicial donde comienza la partición.



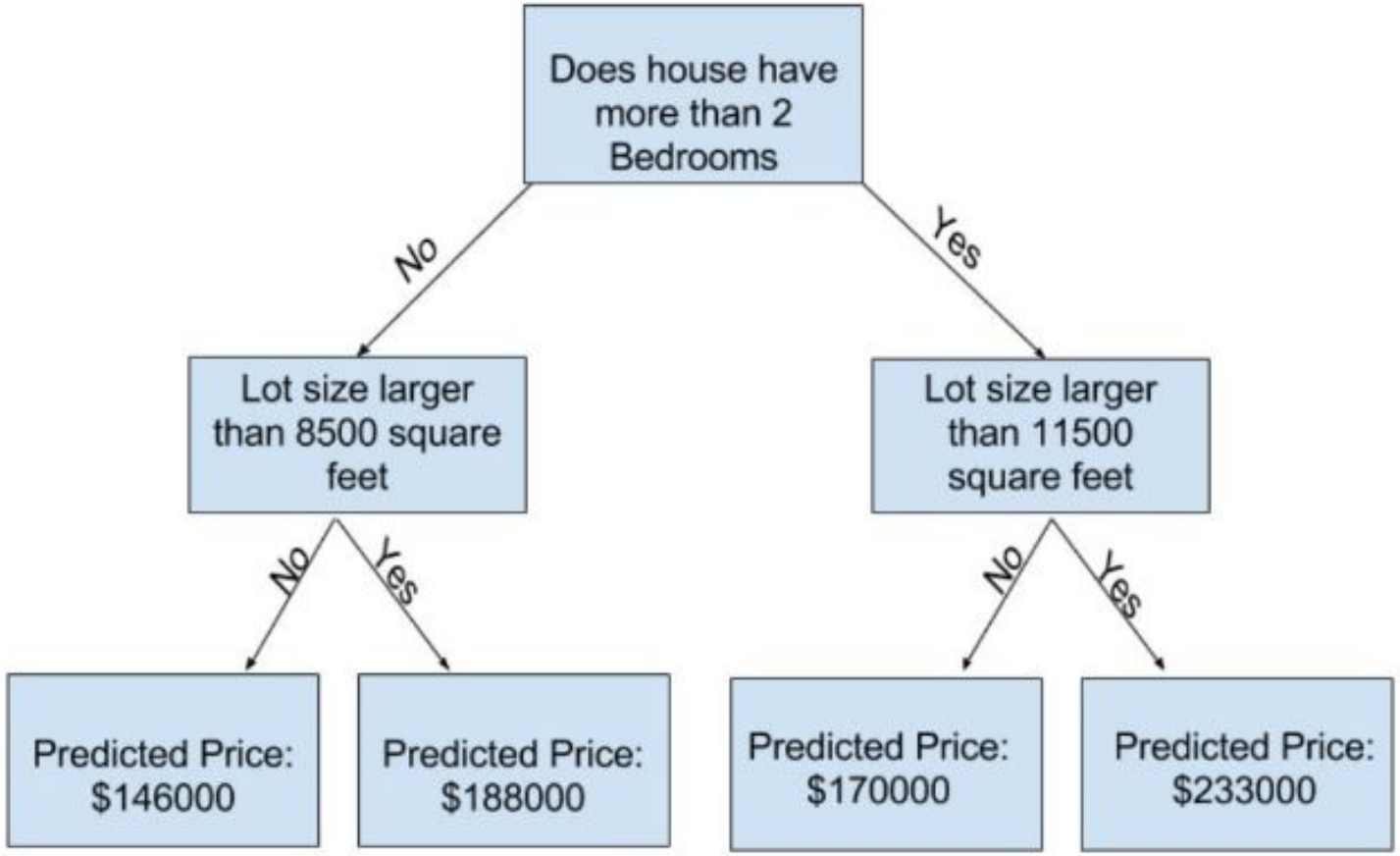
¿Cómo funciona un Árbol de Decisión para Regresión?

- Un árbol de decisión para regresión predice valores continuos.
- Las decisiones se basan en divisiones que minimizan la **varianza** en los datos.



¿Cómo funciona un Árbol de Decisión para Regresión?

- Un árbol de decisión para regresión predice valores continuos.
- Las decisiones se basan en divisiones que minimizan la **varianza** en los datos.



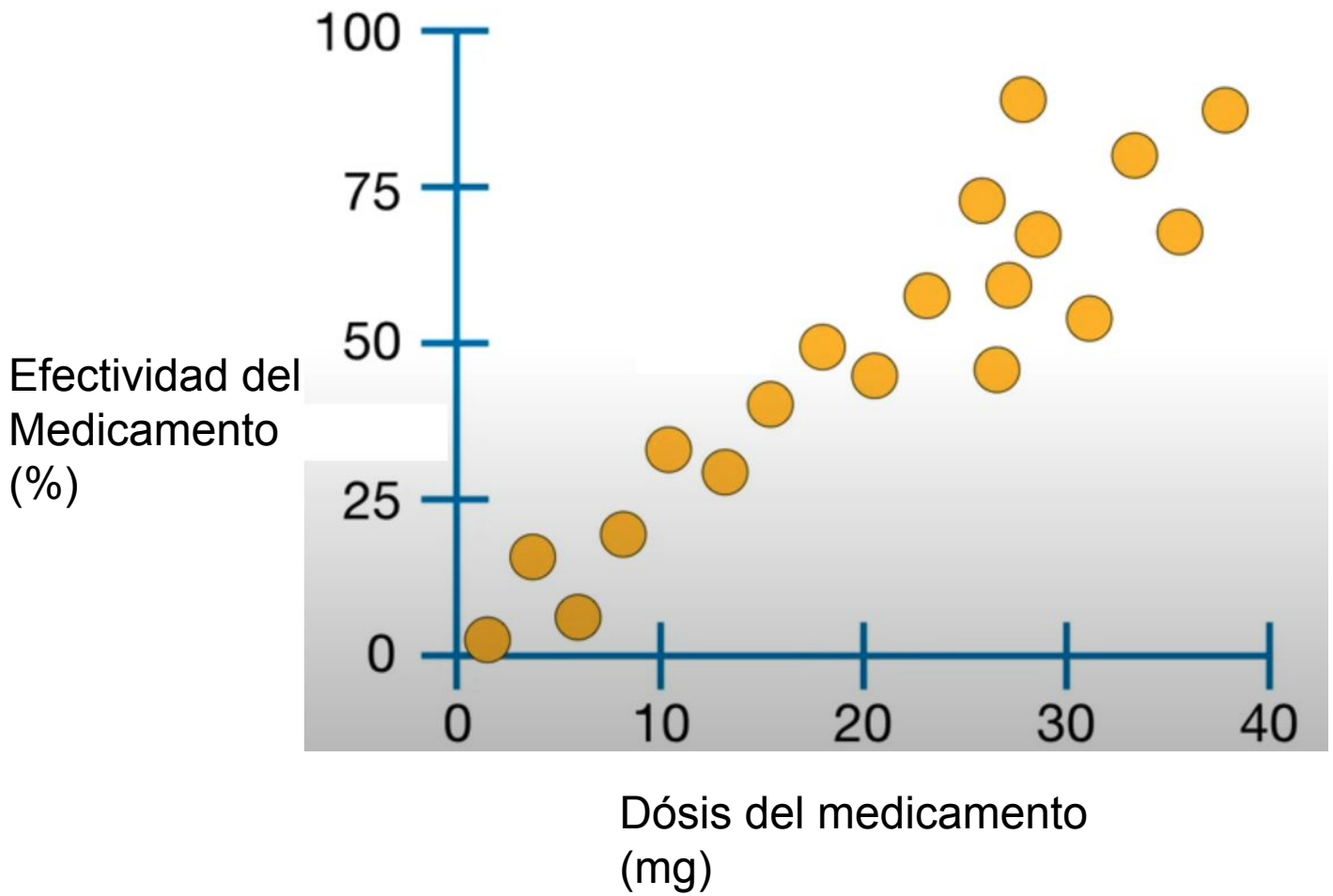
Construcción del Árbol de Decisión

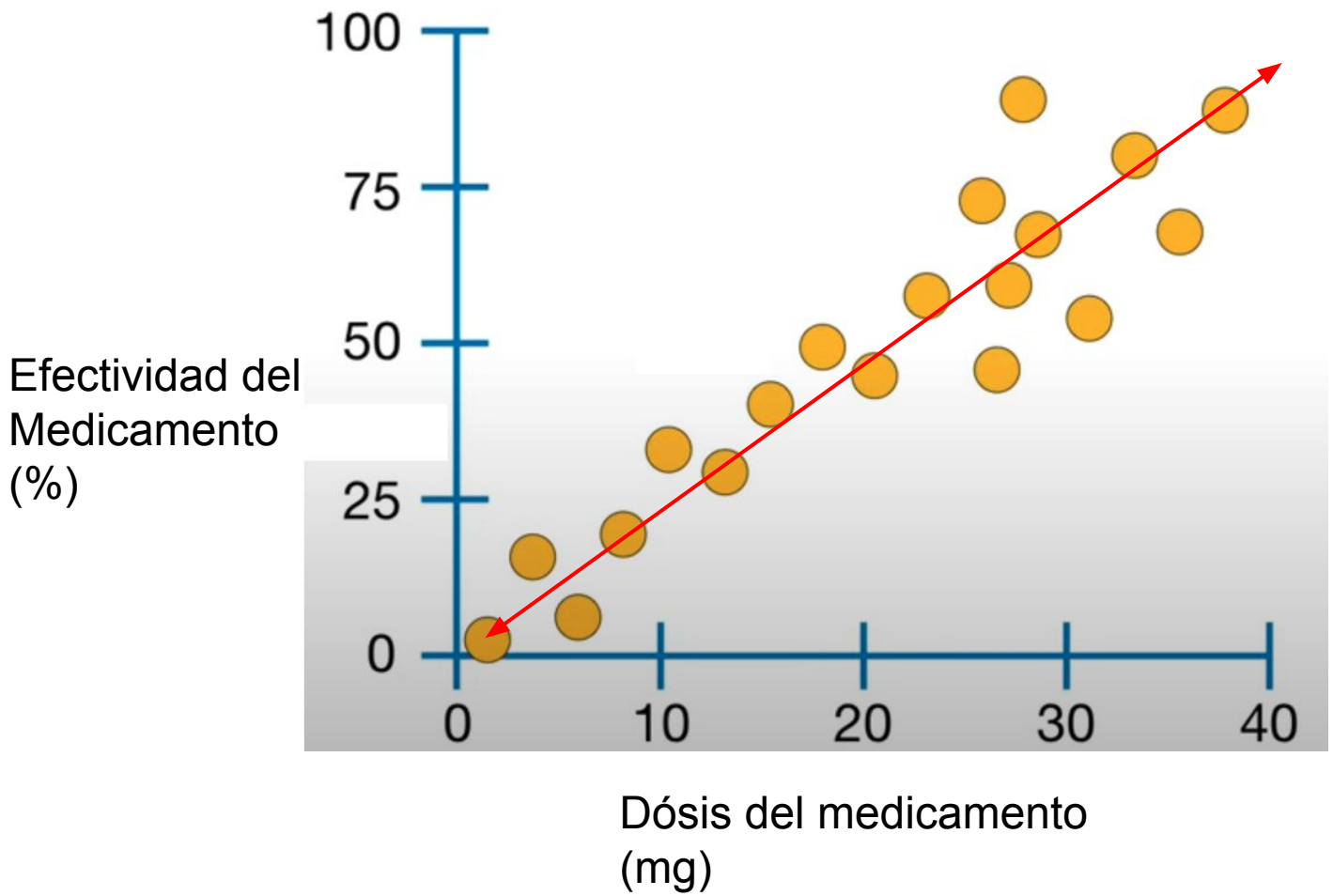
Algoritmo básico:

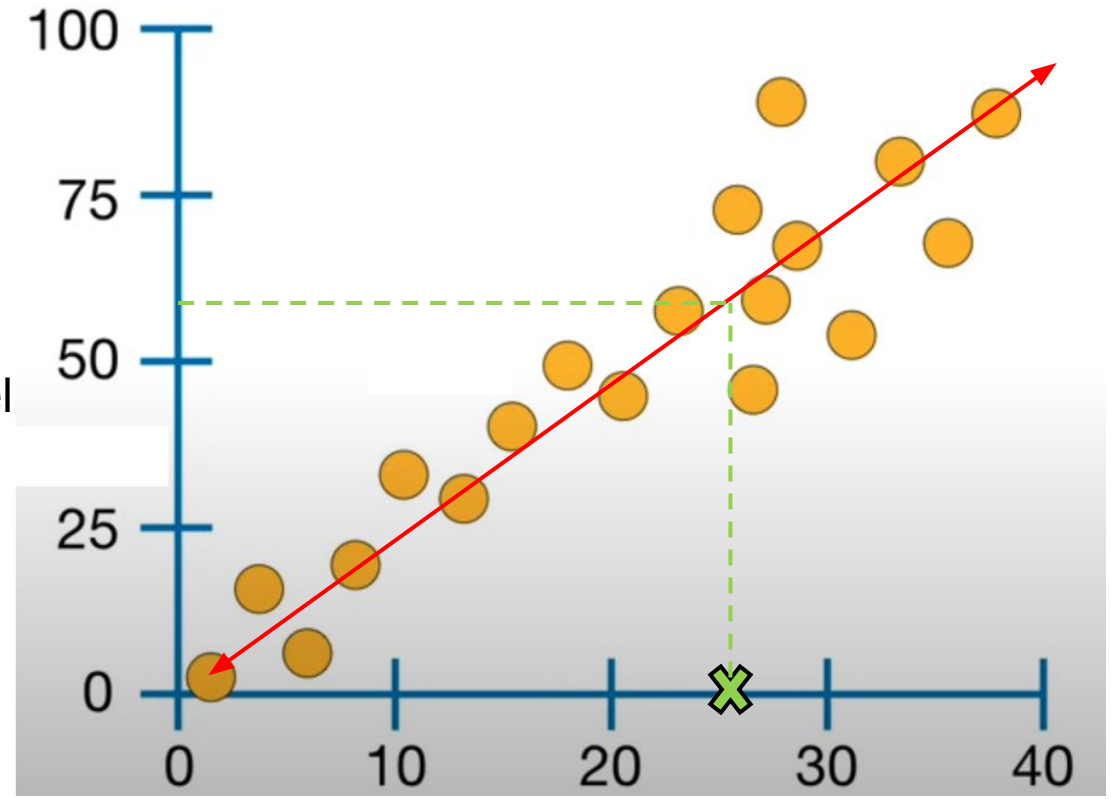
1. Selección de la característica (feature) que minimice la **suma de los errores cuadráticos (SSE)**.

2. División de los datos en subconjuntos.

3. Repetir el proceso hasta cumplir con un criterio de parada (profundidad máxima, número mínimo de muestras, etc.).


Efectividad del Medicamento (%)
100
75
50
25
0
0
10
20
30
40
Dósis del medicamento (mg)


Efectividad del Medicamento (%)
100
75
50
25
0
0
10
20
30
40
Dósis del medicamento (mg)

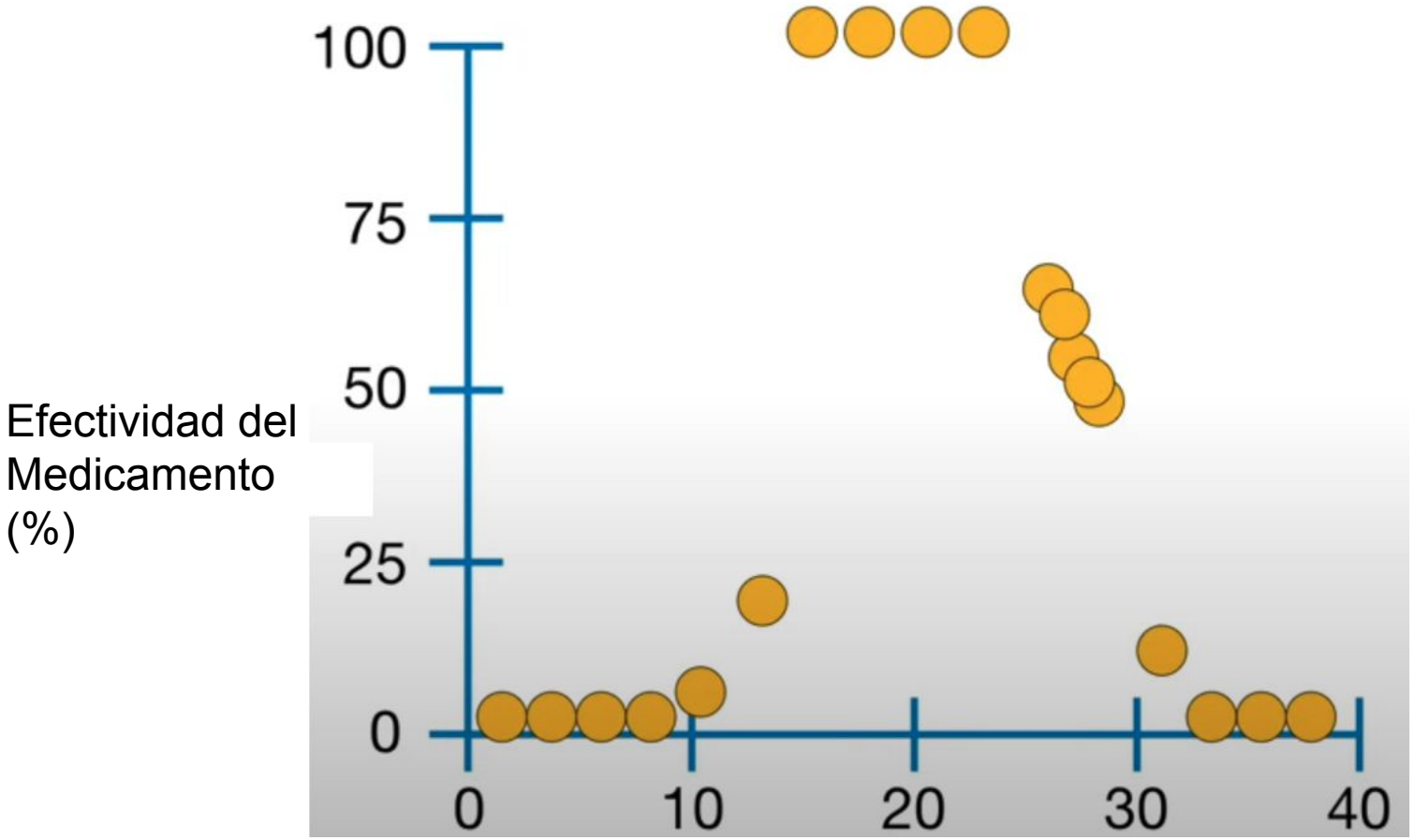


Efectividad del Medicamento (%)

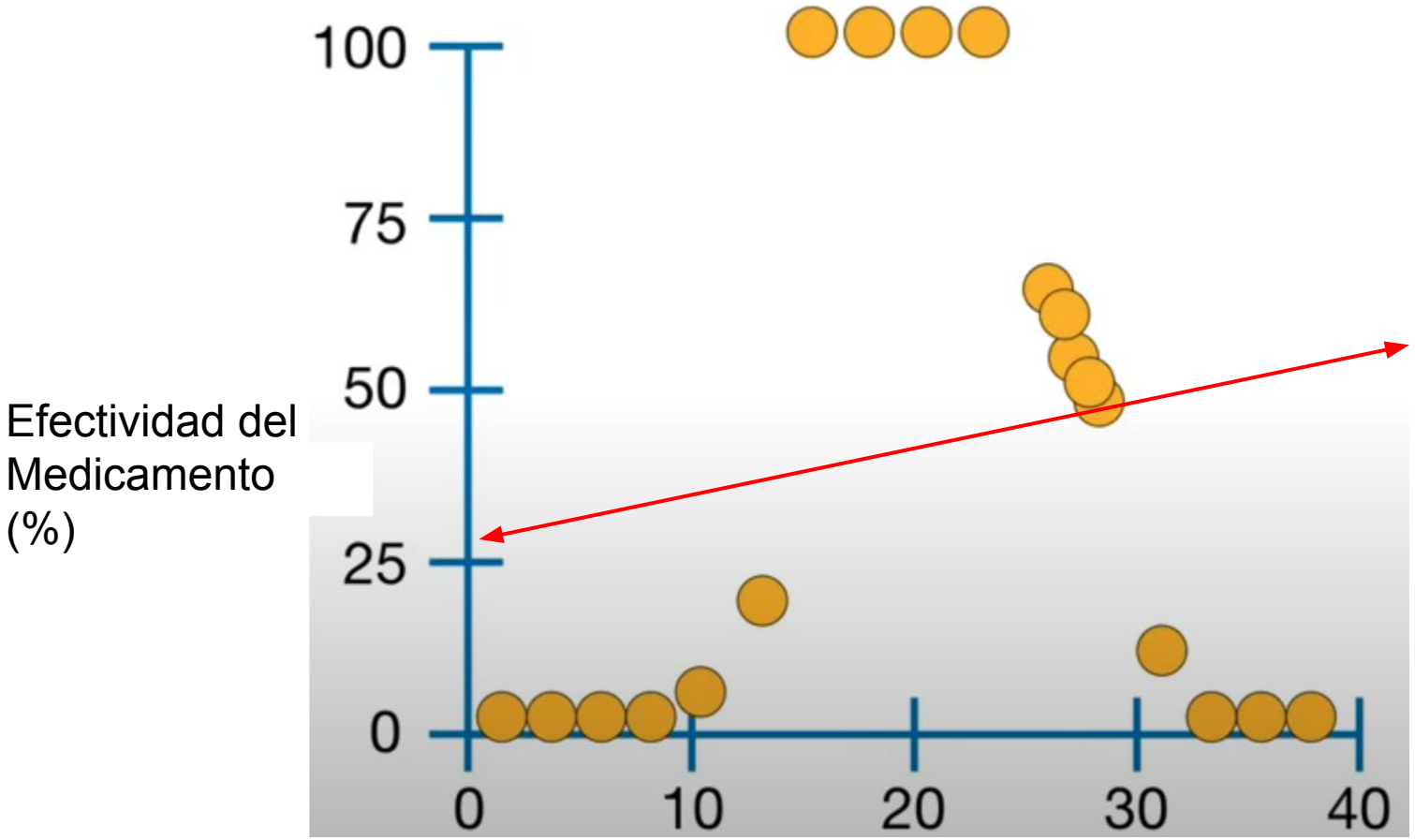
Dósis del medicamento (mg)

Podemos predecir para un nuevo punto dato (e.g. 25 mg) que la efectividad sera 58%.

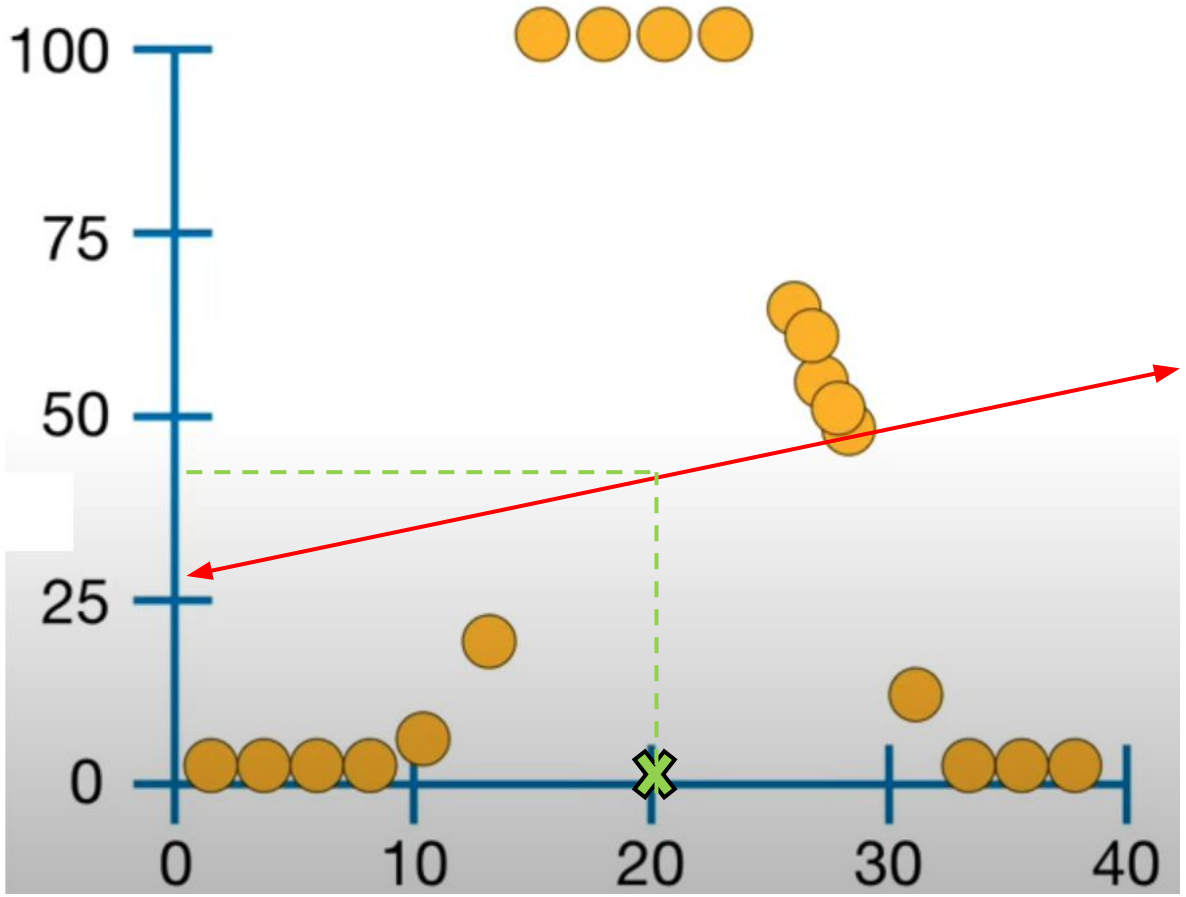
¿Pero qué pasa si los datos no son lineales?



¿Pero qué pasa si los datos no son lineales?



¿Pero qué pasa si los datos no son lineales?

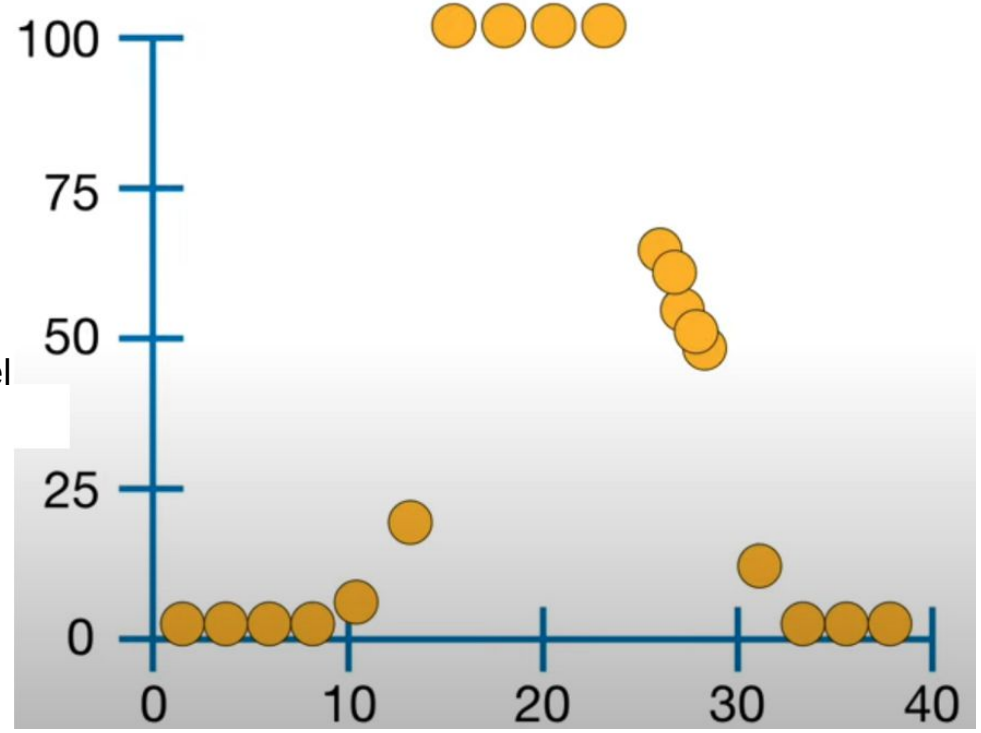


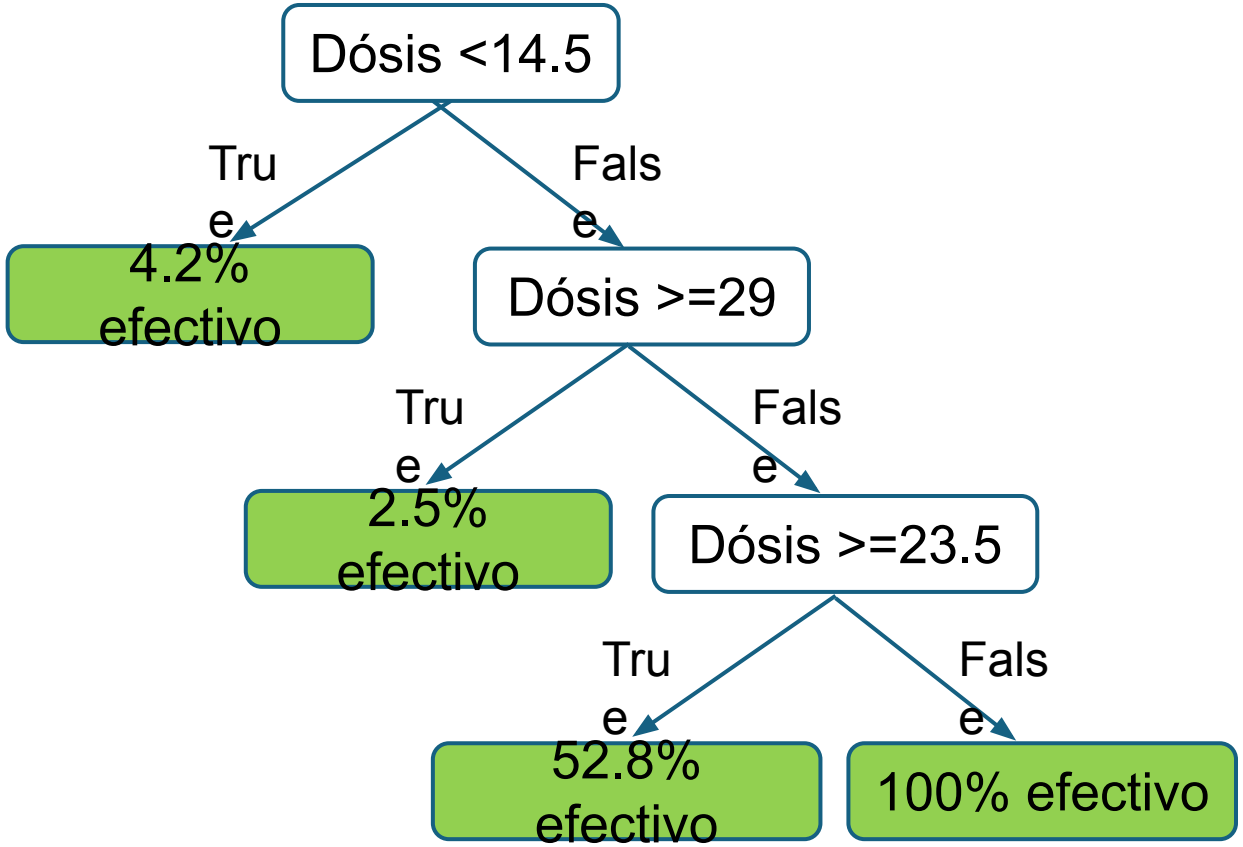
Efectividad del Medicamento (%)

Tendríamos un error muy alto

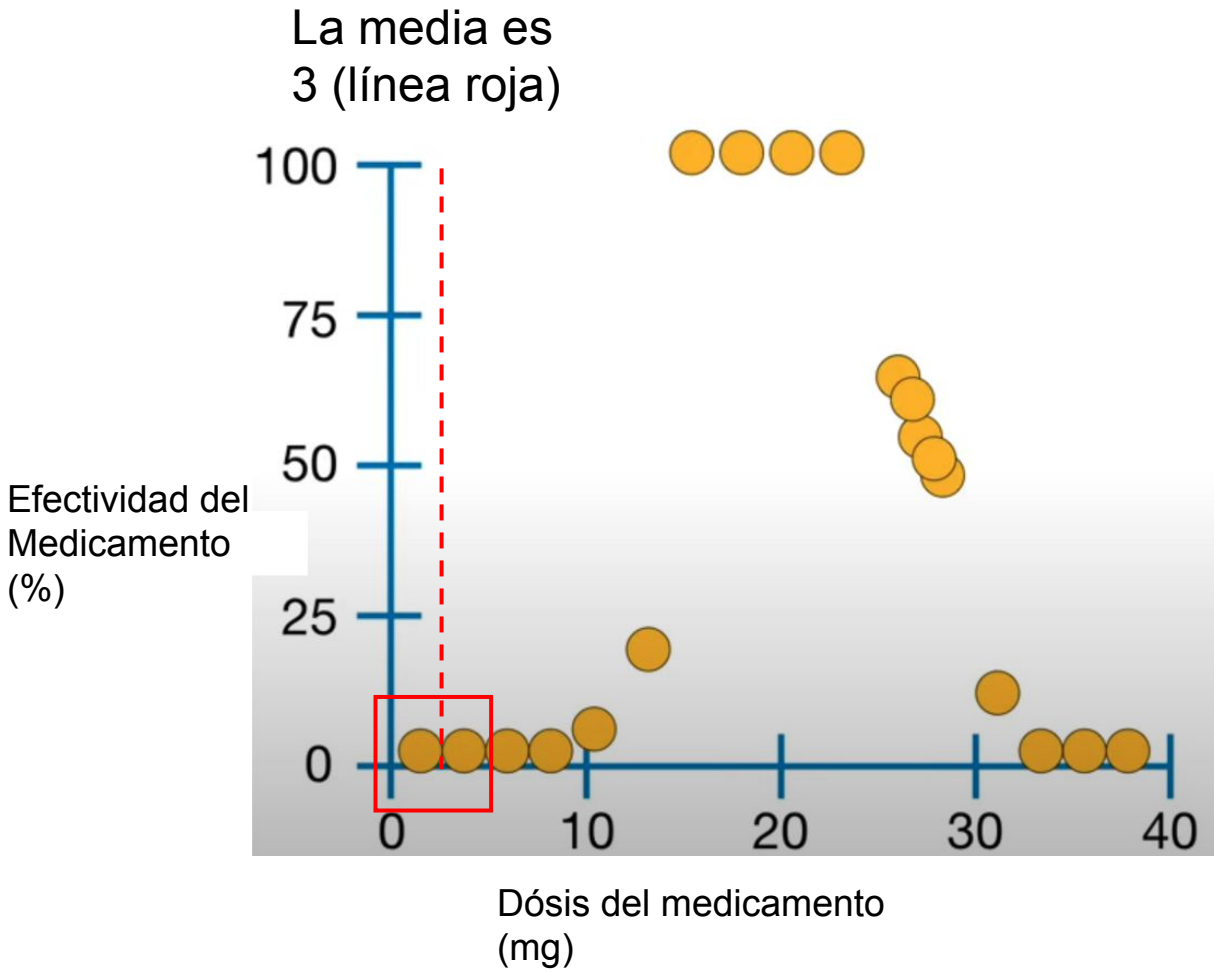
Dósis del medicamento (mg)

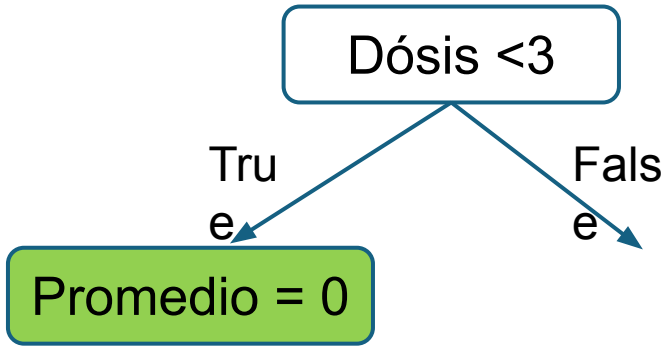
Árboles de decisión (para regresión)



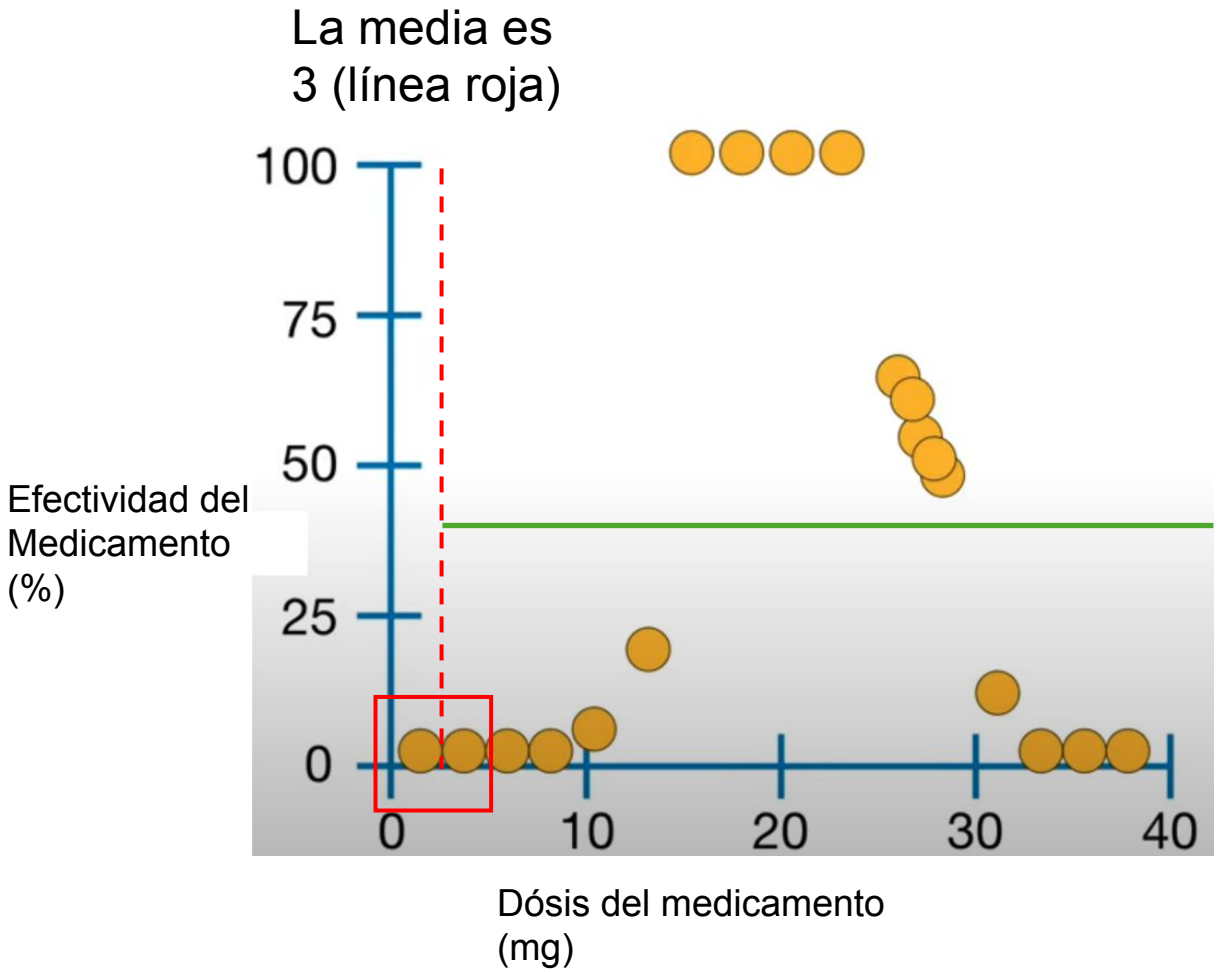


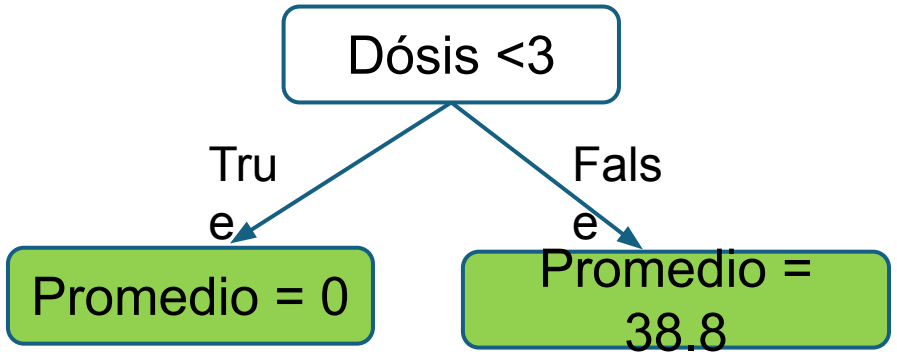
¿Cómo construir el árbol de decisión?



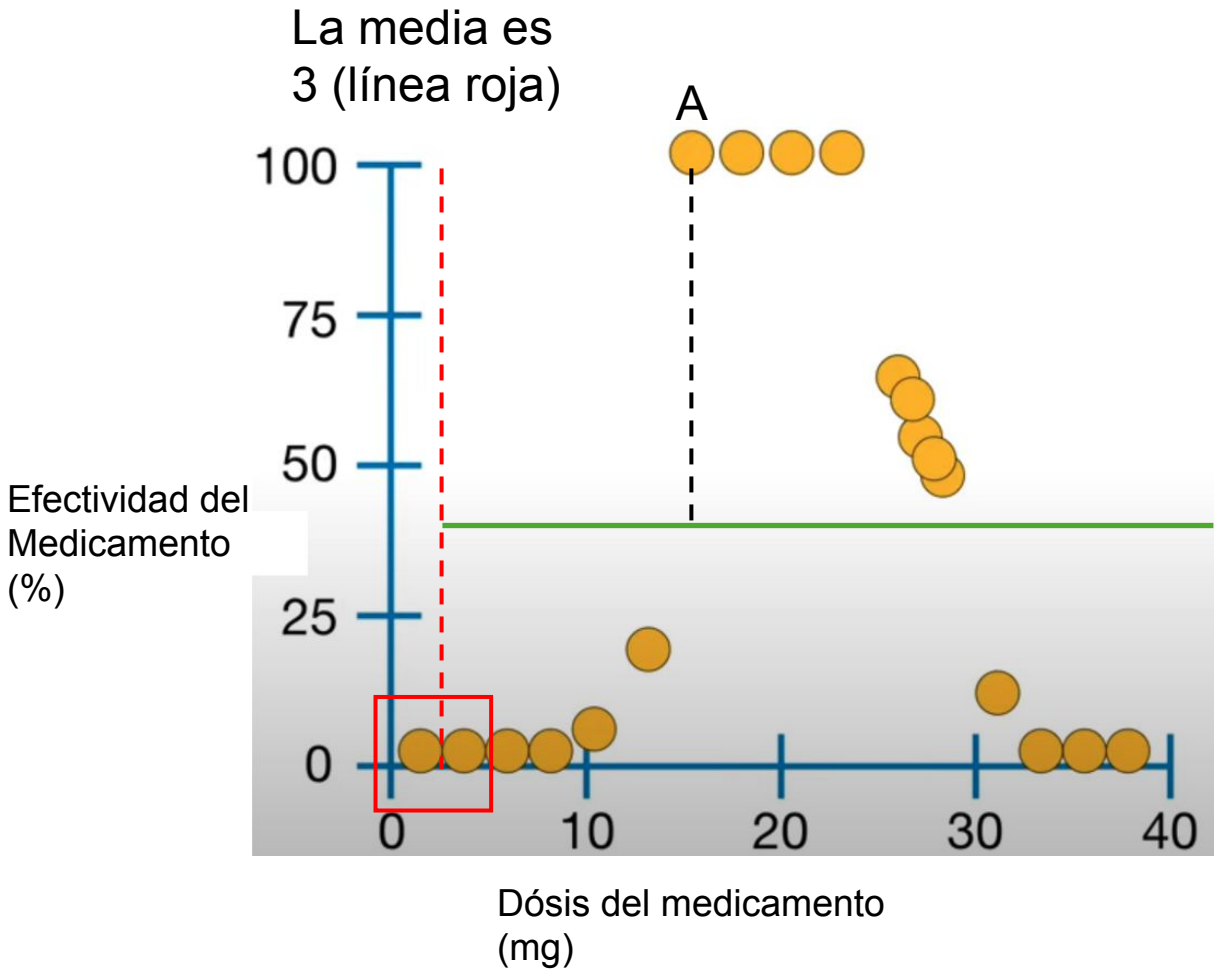


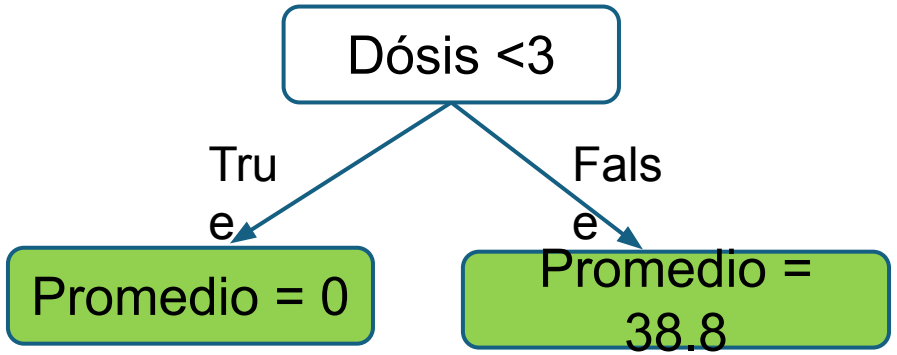
¿Cómo construir el árbol de decisión?





¿Cómo construir el árbol de decisión?

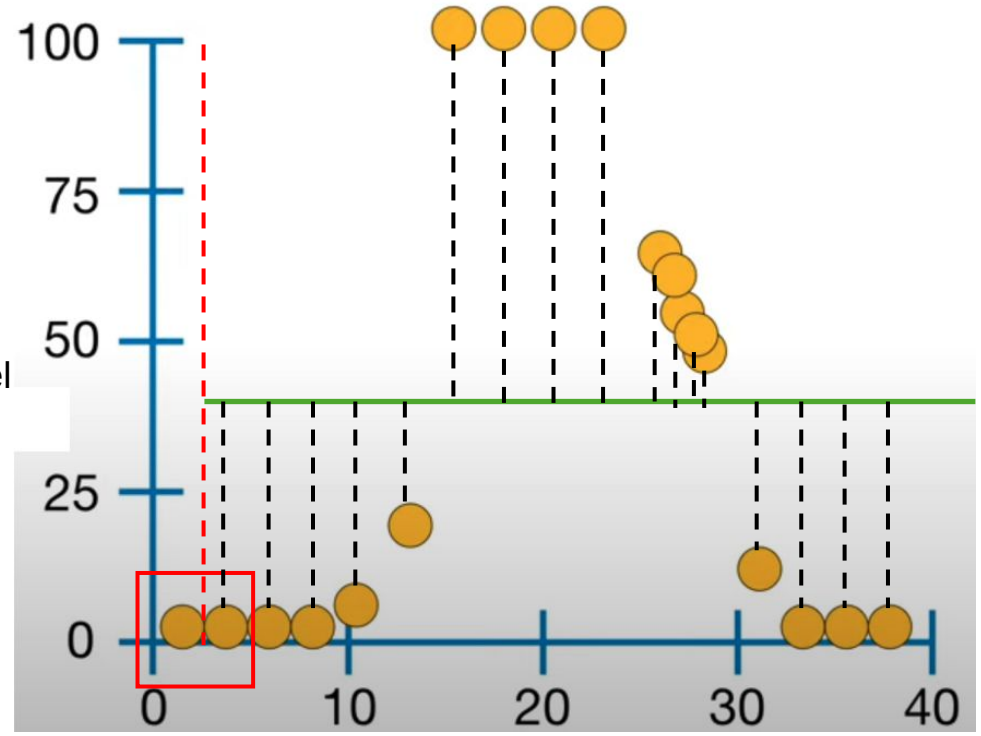




Para el punto A, el árbol de decisión va a predecir una efectividad de 38.8 cuando en realidad es de 100.

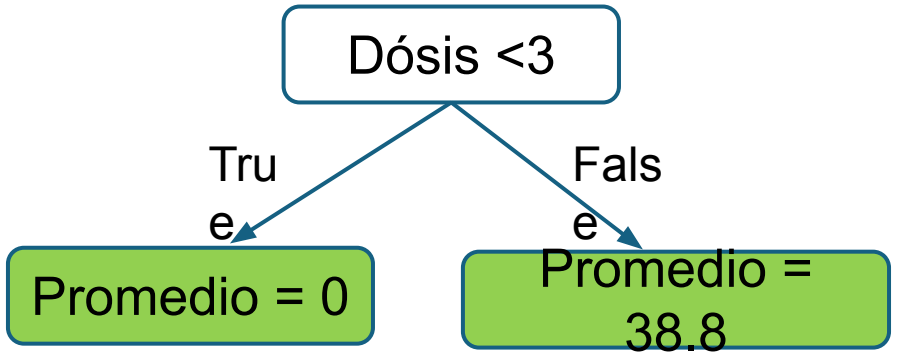
¿Cómo construir el árbol de decisión?

La media es 3 (línea roja)



Efectividad del Medicamento (%)

Dósis del medicamento (mg)



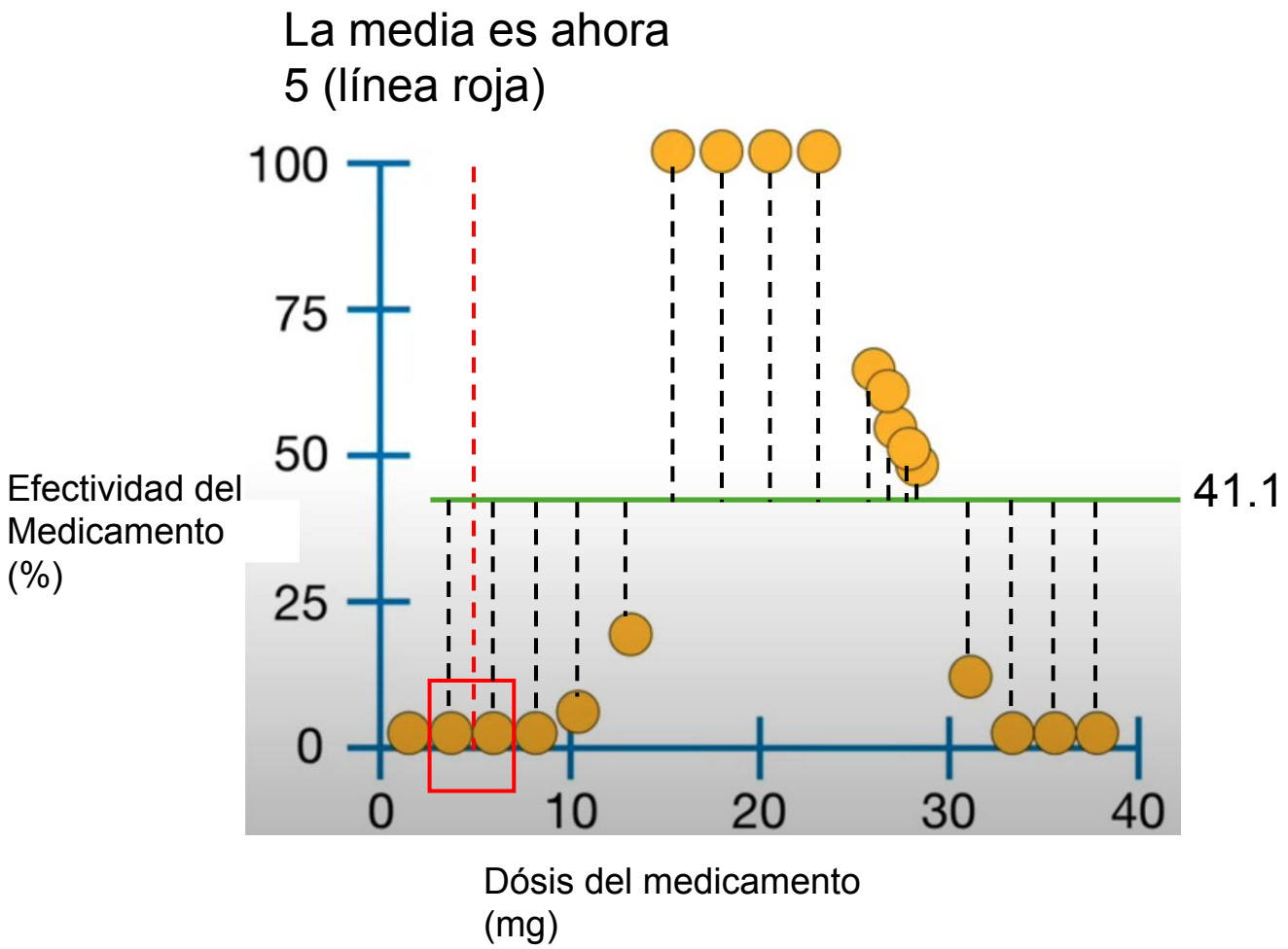
Utilizamos SSE (Sum of Squared errors)

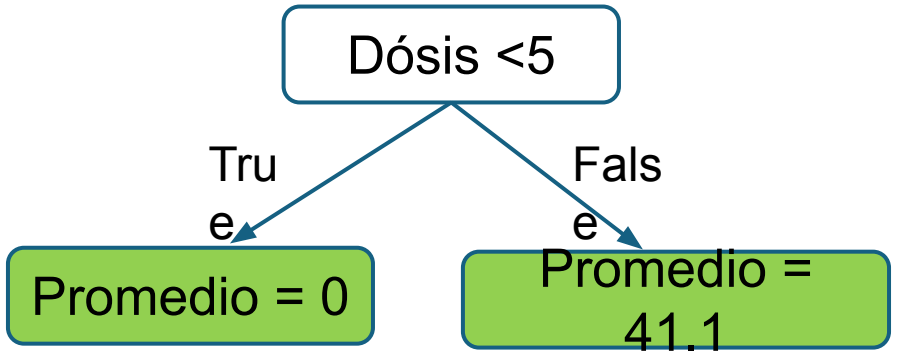
$$SSE = \sum_{i=0}^{n} (y_i - \hat{y}_i)^2$$

$$= (0-0)^2 + (0-38.8)^2 + (0-38.8)^2 \ldots$$

$$= 27468.5$$

¿Cómo construir el árbol de decisión?



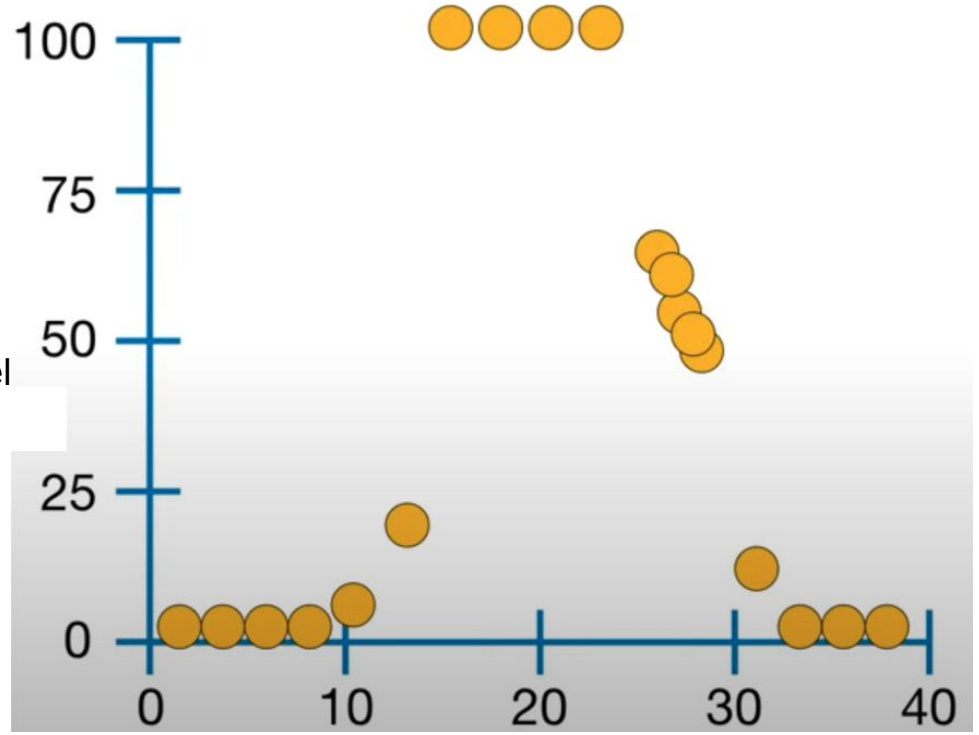


$$SSE = \sum_{i=0}^{n} (y_i - \hat{y}_i)^2$$

$$= (0-0)^2 + (0-0)^2 + (0-41.1)^2 \ldots$$

$$=?$$

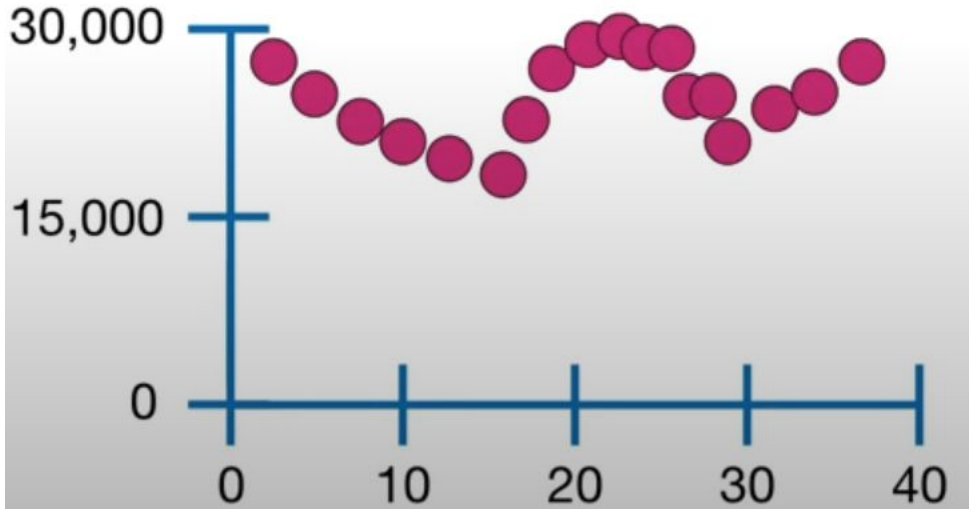
¿Cómo construir el árbol de decisión?



Efectividad del Medicamento (%)

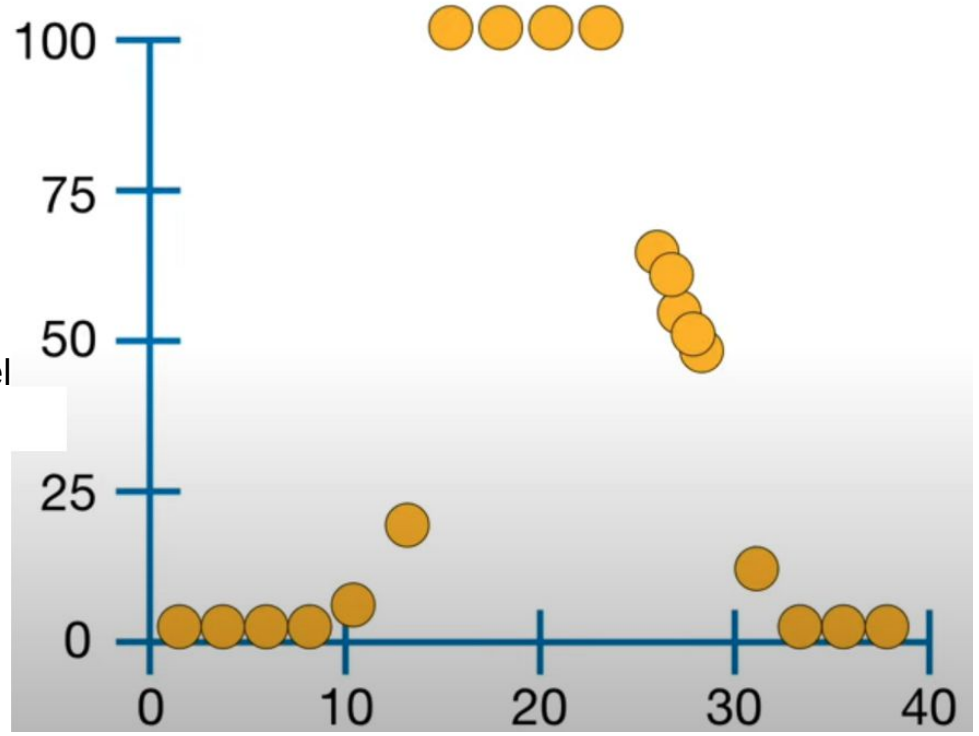
Dósis del medicamento (mg)

SS E



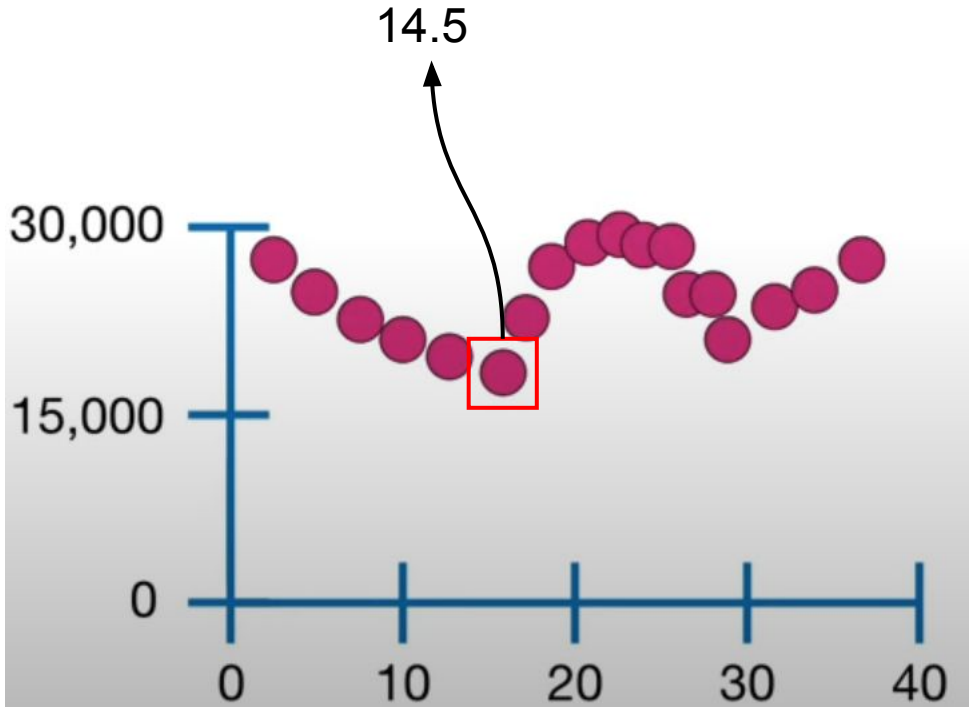
Dósis (Promedio entre dos puntos)

¿Cómo construir el árbol de decisión?



Efectividad del Medicamento (%)

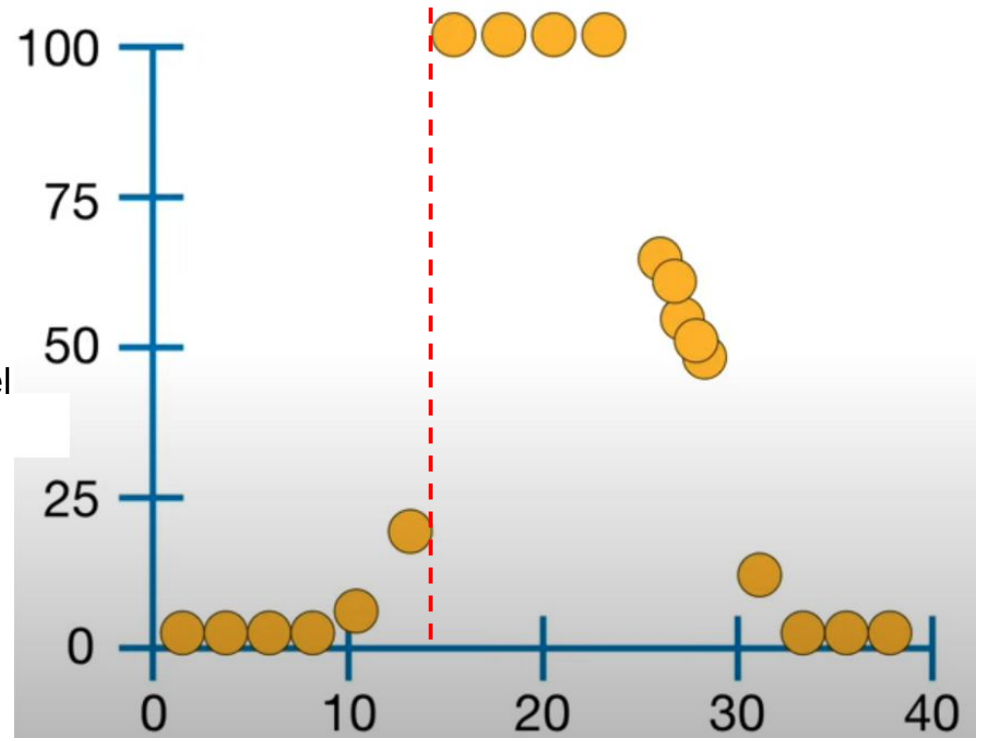
Dósis del medicamento (mg)



SSE

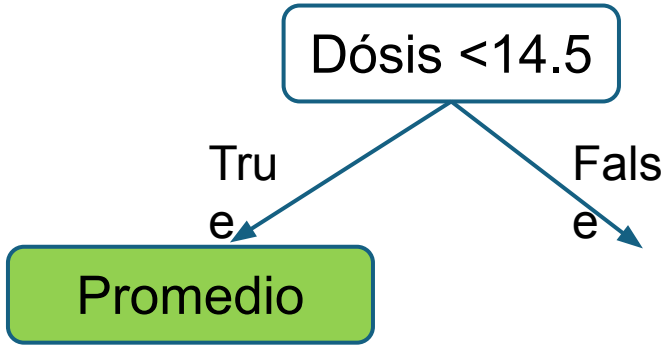
Dósis (Promedio entre dos puntos)

¿Cómo construir el árbol de decisión?

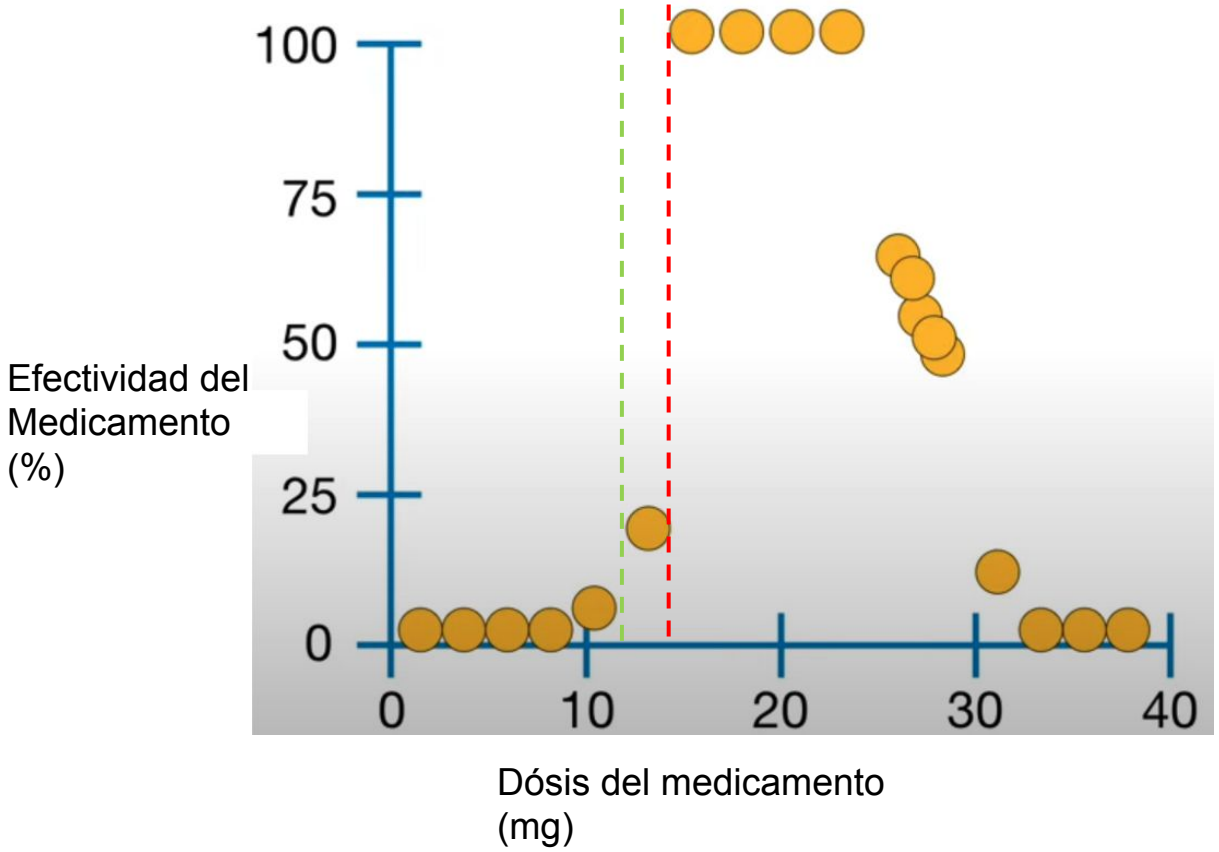


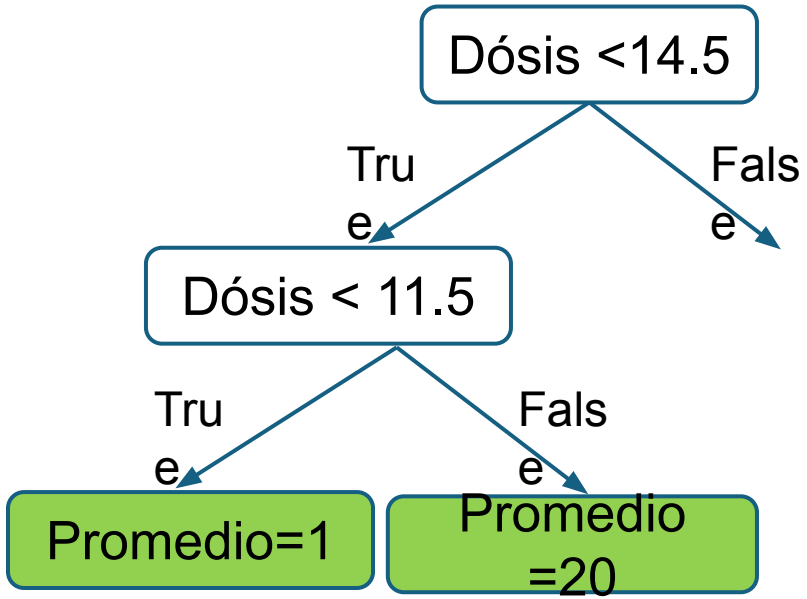
Efectividad del Medicamento (%)

Dósis del medicamento (mg)

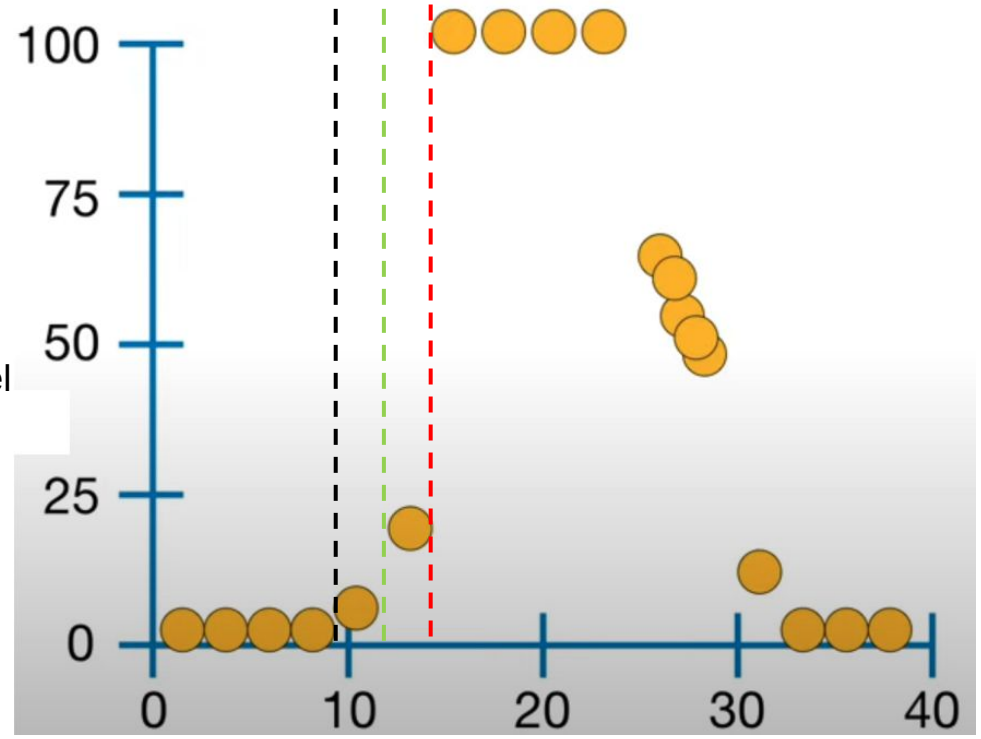


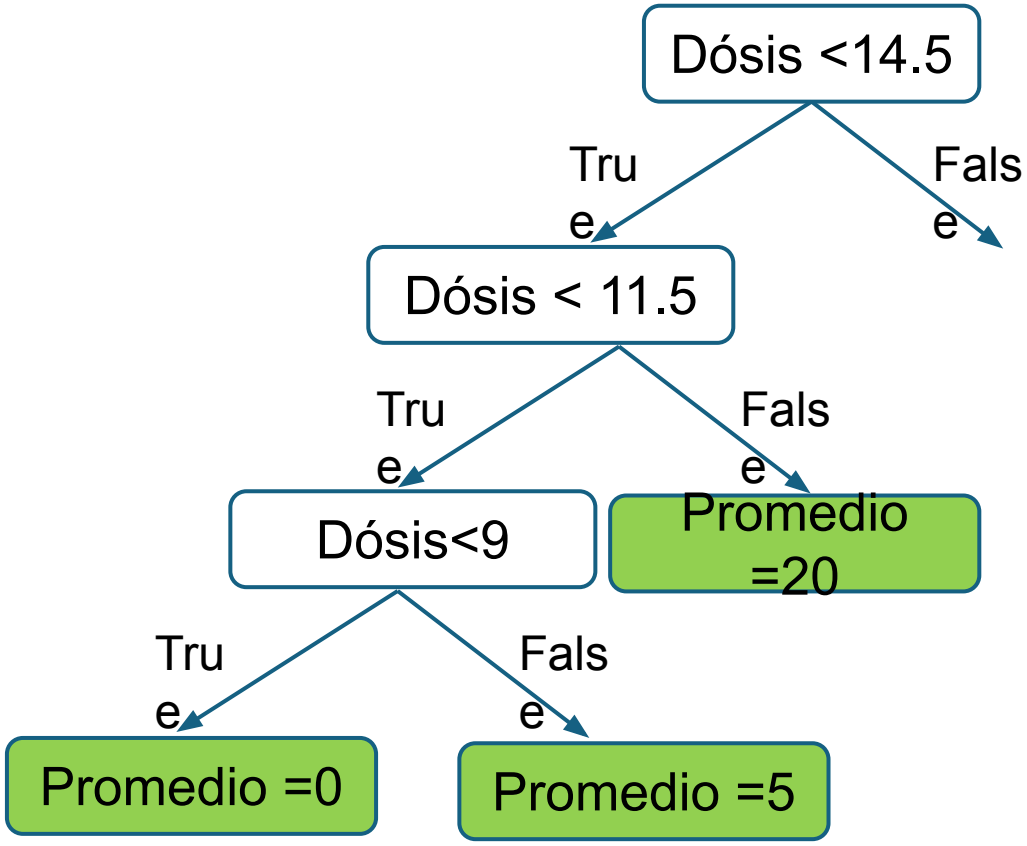
¿Cómo construir el árbol de decisión?





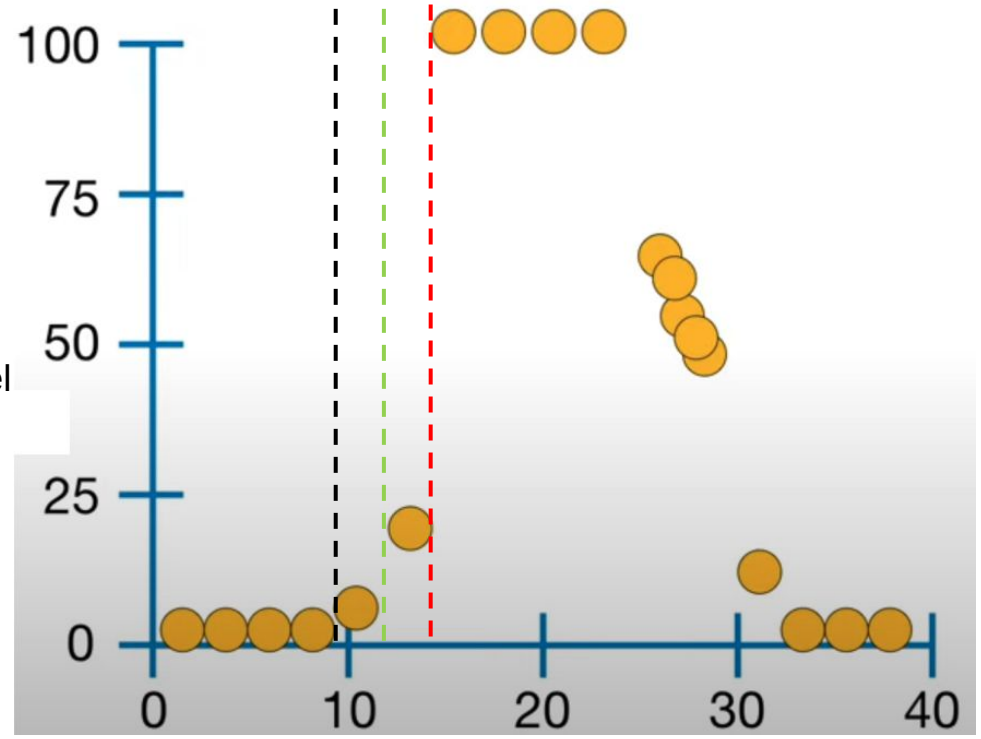
¿Cómo construir el árbol de decisión?

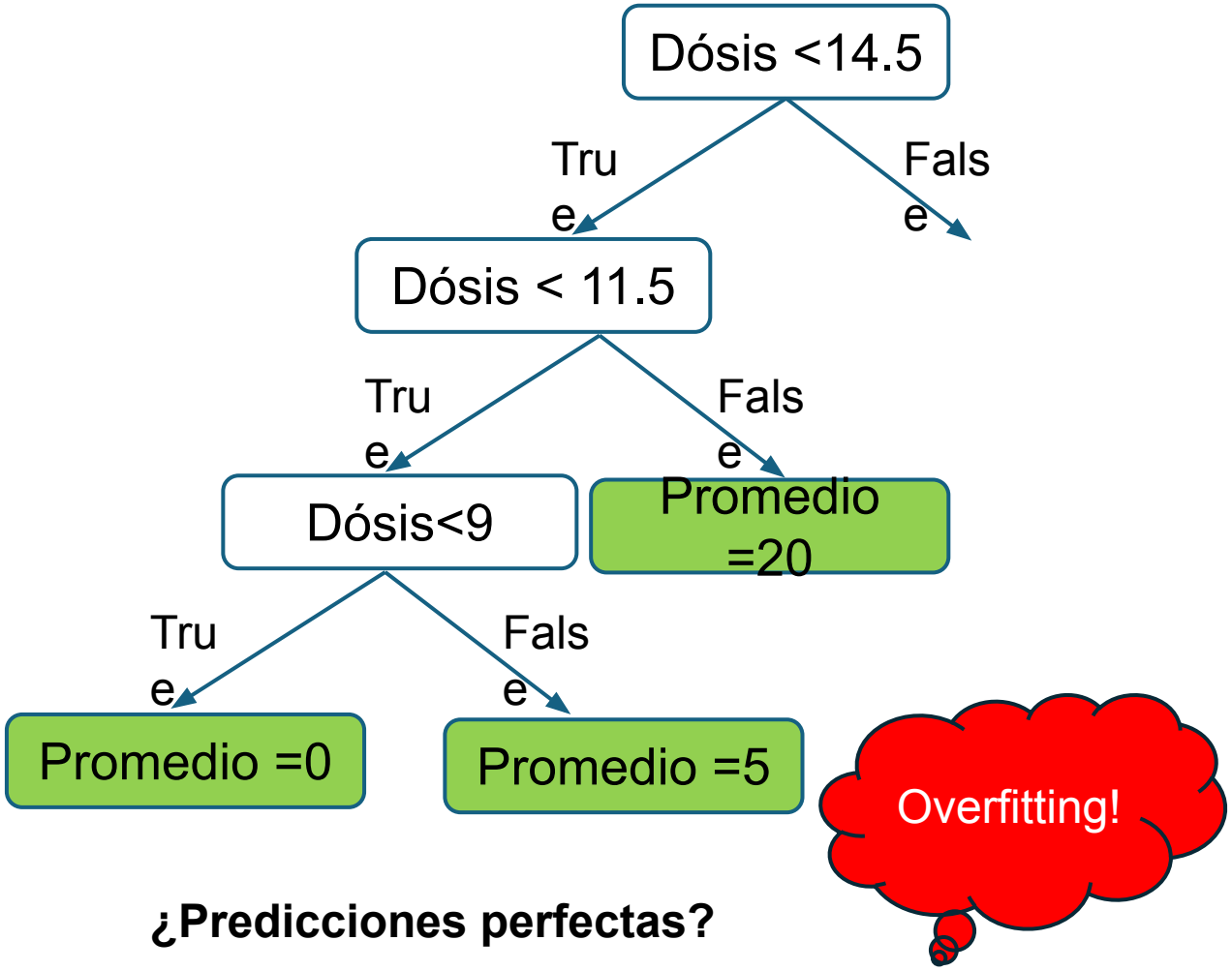




¿Predicciones perfectas?

¿Cómo construir el árbol de decisión?

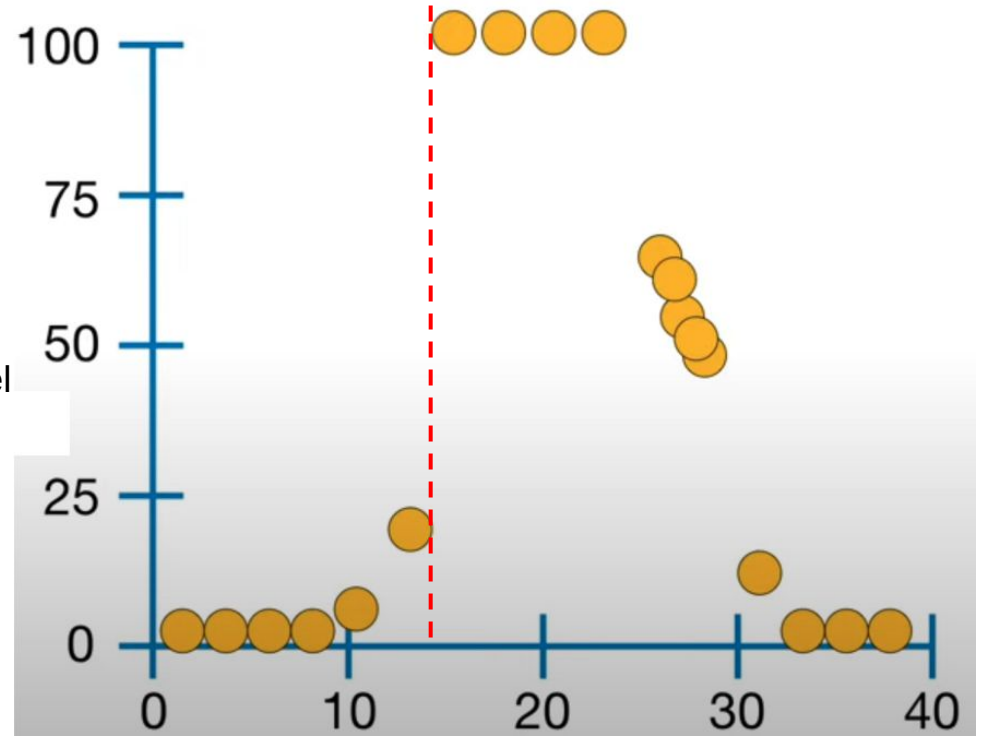


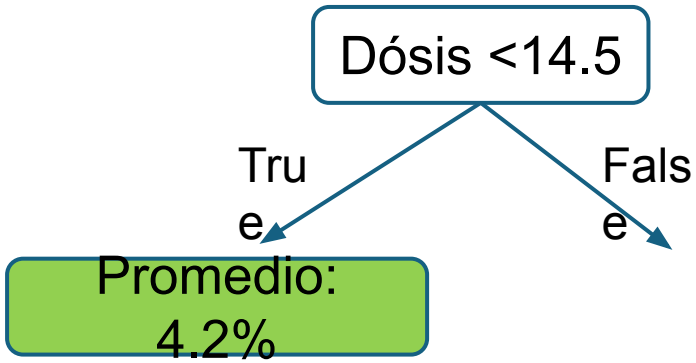


¿Predicciones perfectas?

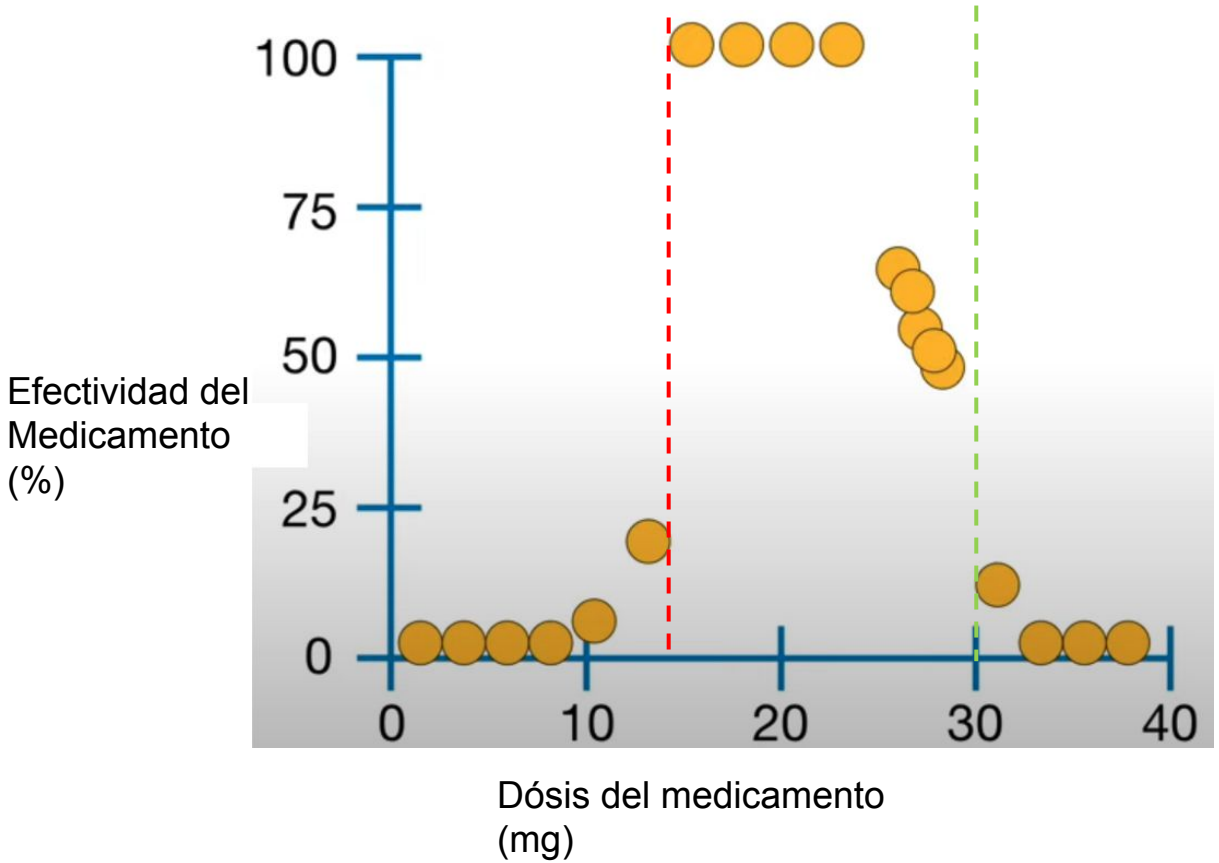
¿Cómo construir el árbol de decisión?

Mínimo número de observaciones = 6

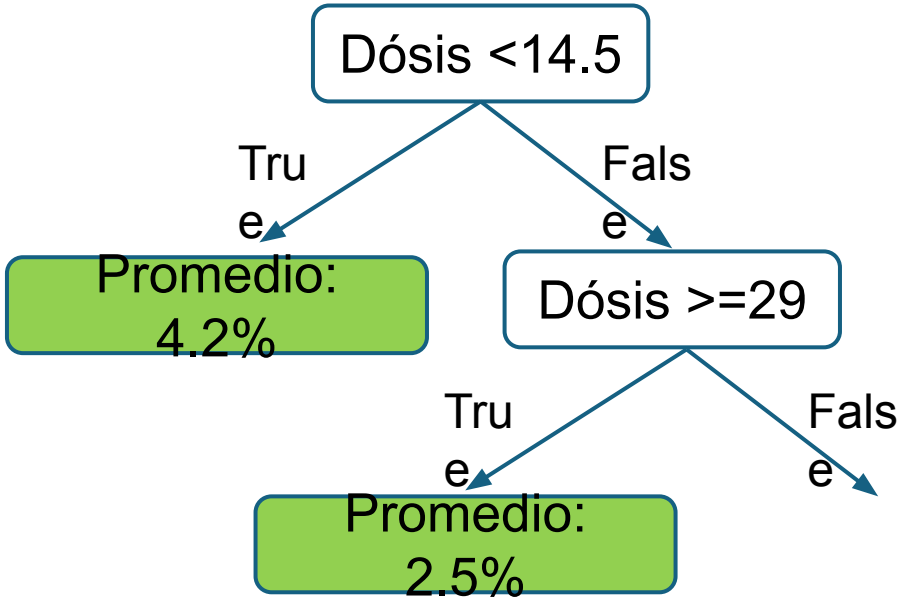




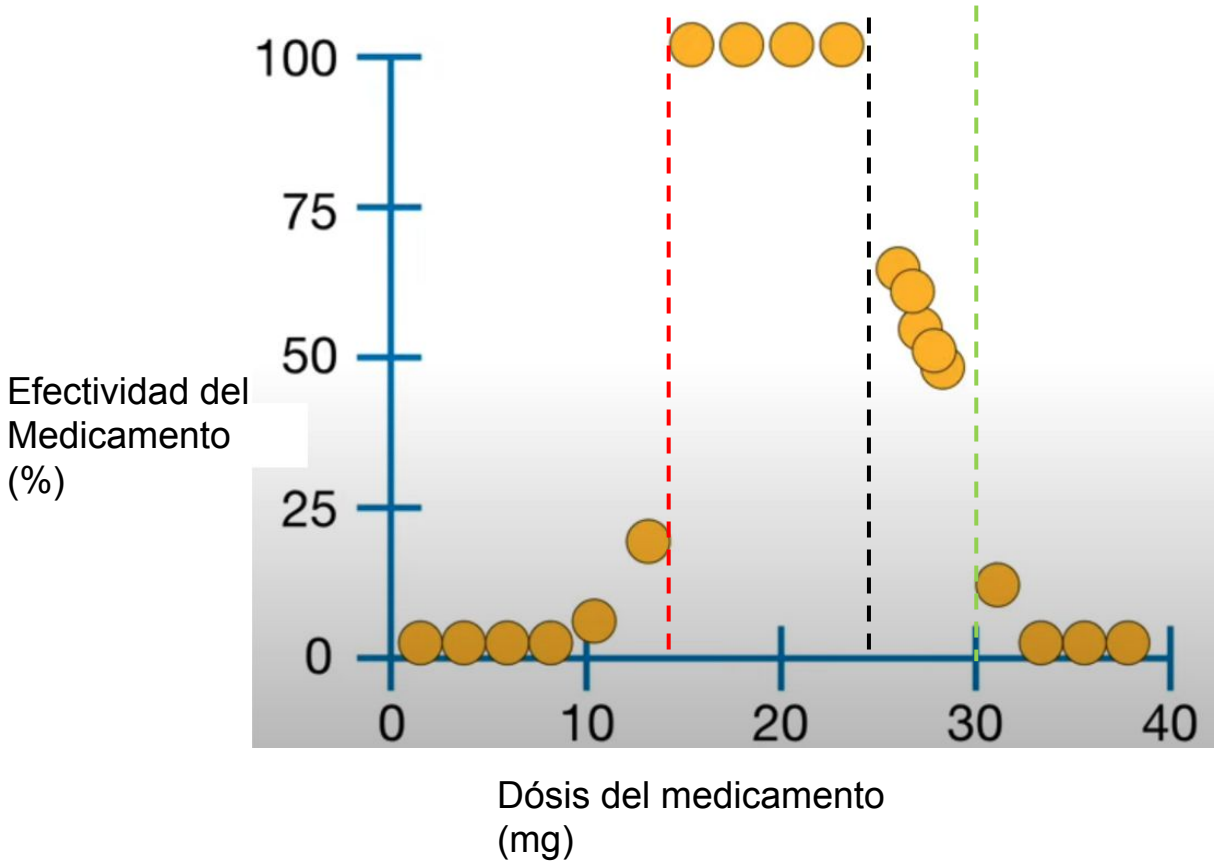
¿Cómo construir el árbol de decisión?



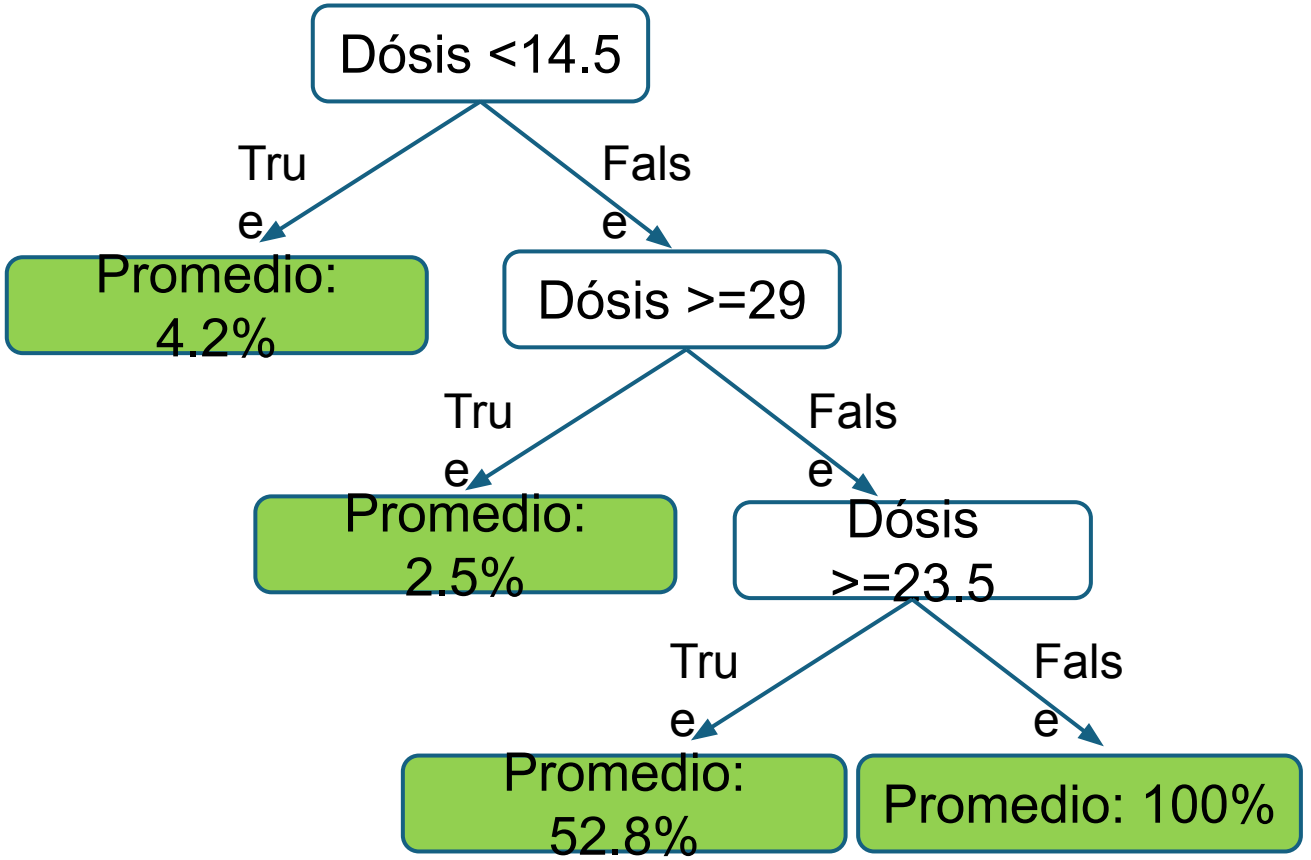
Mínimo número de observaciones = 6



¿Cómo construir el árbol de decisión?



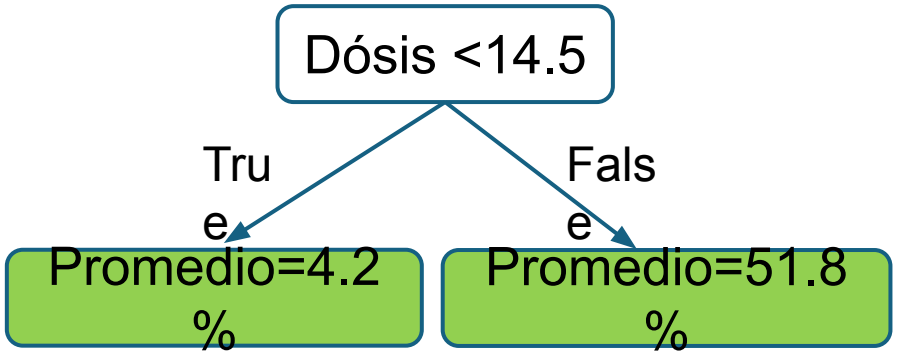
Mínimo número de observaciones = 6



¿Cómo utilizar más features?

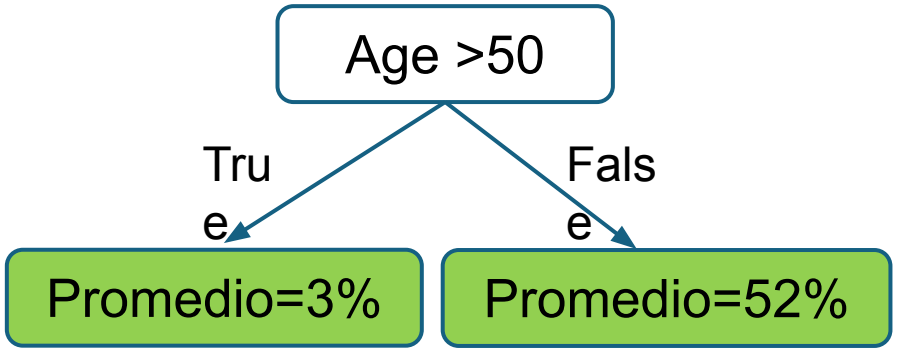
Dósis	Edad	Sex	Efectividad
10	25	Female	98
20	73	Male	0
35	54	Female	6
5	12	Male	44
Etc..	Etc..	Etc..	Etc..

¿Cómo utilizar más features?

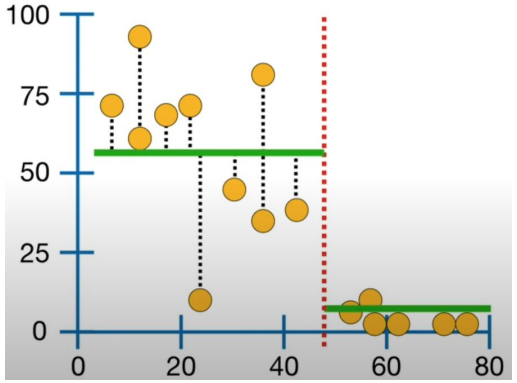


Dósis	Edad	Sex	Efectividad
10	25	Female	98
20	73	Male	0
35	54	Female	6
5	12	Male	44
Etc..	Etc..	Etc..	Etc..

¿Cómo utilizar más features?

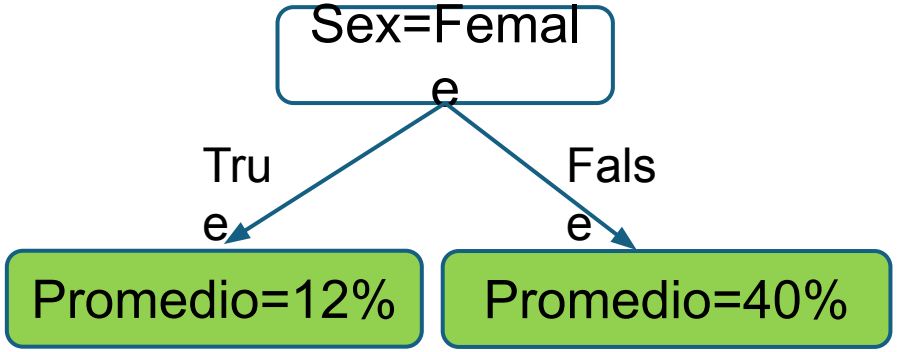


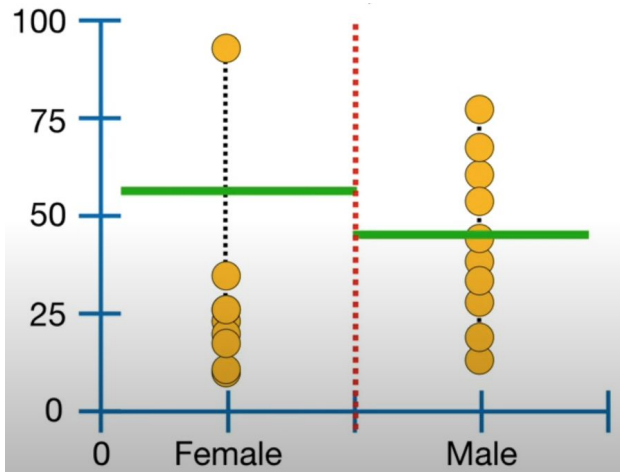
Dósis	Edad	Sex	Efectividad
10	25	Female	98
20	73	Male	0
35	54	Female	6
5	12	Male	44
Etc..	Etc..	Etc..	Etc..



El valor que tenga el mínimo SSE

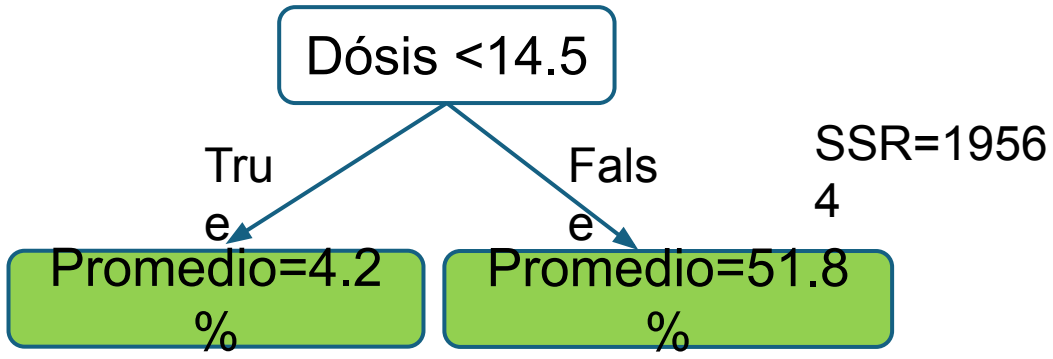
¿Cómo utilizar más features?

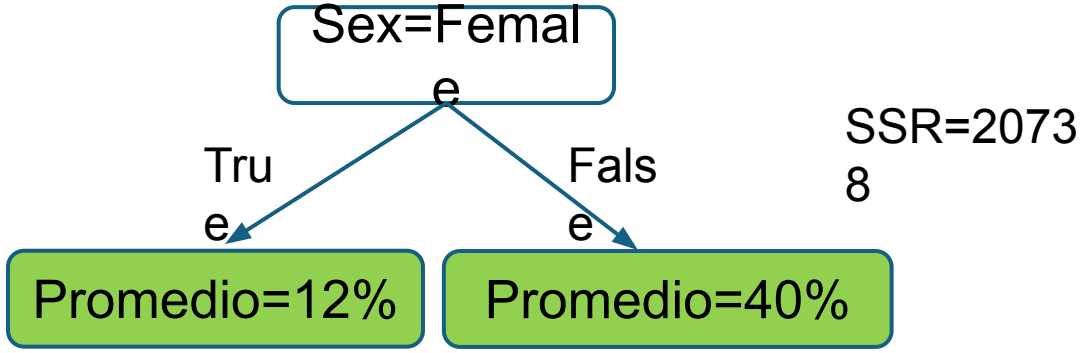


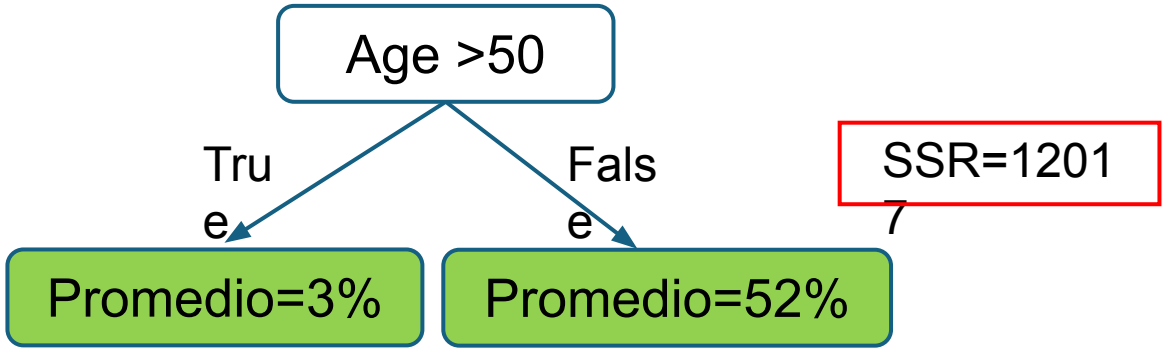


Dósis	Edad	Sex	Efectividad
10	25	Female	98
20	73	Male	0
35	54	Female	6
5	12	Male	44
Etc..	Etc..	Etc..	Etc..

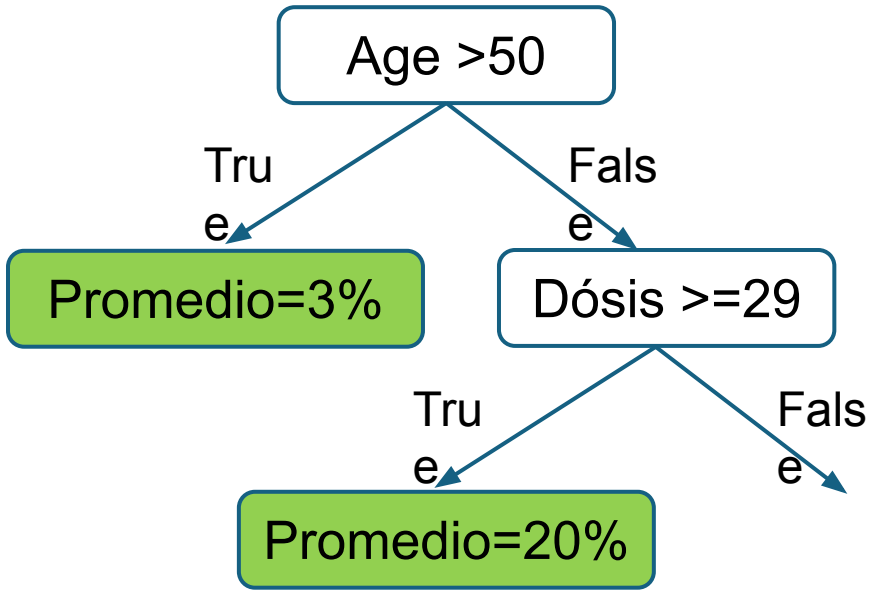
¿Cómo utilizar más features?





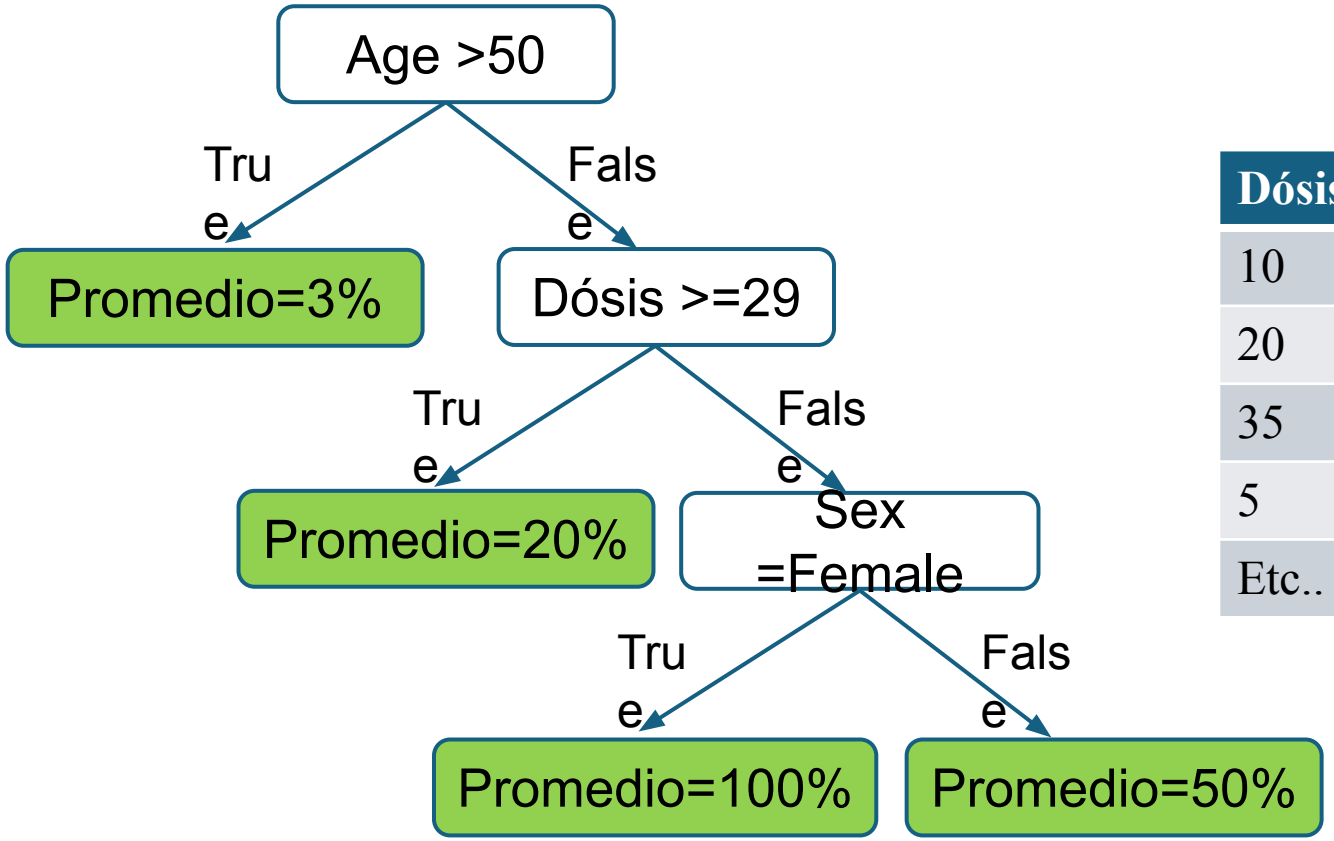


¿Cómo utilizar más features?



Dósis	Edad	Sex	Efectividad
10	25	Female	98
20	73	Male	0
35	54	Female	6
5	12	Male	44
Etc..	Etc..	Etc..	Etc..

¿Cómo utilizar más features?



Dósis	Edad	Sex	Efectividad
10	25	Female	98
20	73	Male	0
35	54	Female	6
5	12	Male	44
Etc..	Etc..	Etc..	Etc..

Podemos seguir creando ramas hasta tener un mínimo número de observaciones.

Ventajas de los Árboles de Decisión

- Fáciles de interpretar: se pueden visualizar fácilmente.
- No requieren preprocesamiento de datos: no es necesario escalar o normalizar.
- Capacidad de manejar datos categóricos y continuos.
- Pueden manejar datos no lineales.

Desventajas de los Árboles de Decisión

- **Propensos al sobreajuste**: Los árboles muy profundos pueden ajustarse demasiado al conjunto de datos de entrenamiento.
- **Sensibles a cambios en los datos**: Pequeñas variaciones pueden generar árboles completamente diferentes.

Evitando el Sobreajuste

- **Podado**: Eliminar ramas innecesarias.
- **Profundidad máxima**: Limitar el número de niveles del árbol.
- **Número mínimo de muestras por hoja**: Establecer un límite mínimo para subdivisiones.

Conclusiones

- Los árboles de decisión son fáciles de interpretar y efectivos para regresión.
- Importancia de evitar el sobreajuste.
- Su rendimiento puede mejorar con técnicas avanzadas como el ensemble learning (combinar varios modelos simples para hacer uno más fuerte)